Gold Punning: studying multistable meaning structures using a systematically collected set of lexical blends

Daniel Kjellander
Introduction and aims of the study

It has been claimed for almost a century that lexical blending is an increasing language phenomenon, especially in the domain of English [Pound 1914; Bryant 1974; Kemmer 2003; López Rúa 2010]. At the same time, several studies report its uncertain, or even marginal, status in linguistic research [Cannon 1986; Kemmer 2003; Brdar-Szabó & Brdar 2008; Fandrych 2008]. Lexical blending would therefore be expected to be an attractive domain of word formation research, as it contains numerous possibilities to explore new ground in this particular field of linguistics. Nonetheless, the well attested complexity of blends, phrased already more than a century ago in Bergström [1906: 27, § 49] as “[h]ere the capriciousness of language operates with great freedom”, seems to have remained an obstacle in efforts to address lexical blending for the larger part of the 20th century (see for instance Bauer [1983]; Cannon [1986]).

It would be no exaggeration to say that the depiction above constitutes a challenge to contemporary word formation research [Cannon 1986; Dressler 2000; Bauer 2012; Gries 2012], and the field has in fact attracted a fair amount of interest in recent years. The increasing number of studies in the last three decades bear witness to this, and the broadened theoretical scope, together with the development of novel tools for data analysis, further emphasize the dynamic character of this field of research.

Despite numerous efforts, the patterns, the distribution, and the motivations of blends have remained unclear in much of previous research [Gries 2004b, 2006]. The difficulty to collect blends in a systematically stringent manner is likely one reason behind this,
and as the title indicates, gold panning is suggested as a metaphor for the hard and
time-consuming work to extract lexical blends from a large body of textual data.
Importantly, punning, as in Gold Punning, is understood here broadly as a strategy to 
achieve effects of wordplay which is an intrinsic characteristic of lexical blending
[Lehrer 1996; Gries 2006; Alm-Arvius 2012; Renner 2015]. Instantiations range therefore
from potentially humorous expressions such as relationsshit (relationship + shit) to more
commonplace items such as glocal (global + local).

Although there are problematic aspects of blend research, the existing literature
addresses constructively a number of topics concerning the nature of blends, and some
of these have resulted in promising steps forward towards more comprehensive, and
more empirically grounded, theories on lexical blending. Among assumptions that have
become more established in terms of general agreement are the tendencies for source
words to display structural (graphemic and phonemic) similarity [Kelly 1998; Gries
2004a; Fábregas & Scalise 2012], the categorical difference between blends and complex
clipping [Bat-El 2006; Gries 2012; Arndt-Lappe & Plag 2013; Beliaeva 2014], and the
distinction between speech error blends arising from conflicting lexical activation and
word formation blends that are the result of intentional lexical processes
[Lehrer 1996, 1998, 2007; Beliaeva 2014]. Viewpoints range from attitudes assuming that seriality in any form is incompatible with the concept of
lexical blending [Fradin 2015] to approaches including most items that display some
sort of truncation and fusion of form [Plag 2003; Lehrer 2007]. To address such complex
issues of categorization, it has been suggested that a prototype theory perspective is
appropriate to account for the complexity of lexical blending [Brdar-Szabó & Brdar
2008; Renner 2015]. Because prototype theory is commonly applied in a Cognitive
Linguistics (CL) framework [Taylor 2003; Geeraerts 2006c; Ungerer & Schmid 2006], it
seems natural that several accounts have put forth CL as an ideal theoretical outset
when studying blends [Kemmer 2003; Gries 2006; Silaški & Đurović 2013].

Further examples of issues that remain largely unresolved are the presumed semantic
relatedness between the lexemes that produce the constituent segments, traditionally
called the source words (SW), of a blend [Algeo 1977; Brdar-Szabó & Brdar 2008] and the
prototypicality of certain formation patterns such as distribution of truncation
strategies and source word overlap [Renner et al. 2012; Arndt-Lappe & Plag 2013].

Up until the 1990s, most accounts were concerned with classification and
categorization of lexical blends [Kemmer 2003], and contemporary work has thus
shifted the focus of attention towards the structure of blends, and efforts have been
made to provide statistical evidence to arrive at more empirically robust conclusions
[Gries 2004b, 2006]. On good grounds, it is thus reasonable to say that the knowledge
about lexical blends has increased substantially in the last thirty years. It is, however,
also apparent that many questions await further and more elaborate answers, which is especially true as to the semantics of blends and the motivations for their existence [Gries 2006; Bauer 2012; Gries 2012].

A recurring issue related to a large part of previous studies is how to collect data in a systematic fashion so as to ensure that the findings and conclusions are, if not generally applicable, at least allow replication and representativity within stipulated limitations. As will be discussed below, the present study departs from the standpoint that this is indeed a crucial concern in contemporary blend research.

The aim of the current study is two-fold. First, a methodological approach is suggested that enables quantitative and qualitative analyses based on a systematically collected dataset. The importance of developing systematic data retrieval procedures in blend research can hardly be overestimated, which is why this aim has been given slightly more prominence in this account. Central issues concern the representativity of the data, their relevance and potential for generalizability, and delimitations related to pragmatic aspects. For instance, if the data originate from a specific jargon, it cannot be taken for granted that usage patterns are automatically transferable to other domains of language. Therefore, the source of the data, involving aspects such as temporal scope, genre, mode, and culture, is crucial.

Second, there are several reports on the influence from general cognitive mechanisms on the formation and use of lexical blends [Kelly 1998; Kemmer 2003; Gries 2004b; Kjellander 2018]. Recent findings show a number of observable structural tendencies in blends [Gries 2012; Arndt-Lappe & Plag 2013; Beliaeva 2014; Juhasz et al. 2016], but indications of influence from more general cognitive processes also suggest that a wider scope is necessary to understand lexical blends better [Kelly 1998; Kemmer 2003; Gries 2004b; Fandrych 2008; Bauer 2012; Ronneberger-Sibold 2012]. The second aim of the study is therefore to investigate a particular semantic aspect of the blends in the data, namely how lexical ambiguity is exploited as a means to achieve certain communicative effects. A central research question that was applied for this purpose is phrased: how is lexical ambiguity realized in lexical blends in a contemporary US web news context?

Ambiguity in this context is understood as lexical structures allowing, or even driving, meaning construal involving unresolved semantic conflicts. While ambiguity is a language phenomenon sometimes evoking interest and humorous reactions, but perhaps more commonly thought of as a communicative problem, it is investigated what potential purposes it might serve in lexical blending. The theoretical outset, a Cognitive Linguistics (CL) perspective, means that meaning construal is seen as a basis for all semantic operations concerned with the conceptualization of utterances and comprehension [Langacker 2008]. Meaning construal is furthermore understood as a dynamic process, in which multiple modalities come into play [Geeraerts 2006b], and ambiguity influencing meaning construal can thus be described as an inherently cognitive mechanism drawing on general cognitive abilities. In short, this mechanism is understood as a cognitive constraint influencing the formation and use of lexical blends (cf. Gries [2004b]; Kjellander [2018]).

The aims of the investigation are pursued within the context of US web news between January 2010 and March 2018. Because the intention of the study is to study patterns of present-day blend usage, a large contemporary database is required, which is why the entire corpus up to the time of the data download was chosen. This means that 99
For reasons of space, only common nouns as potential sources of lexical blends are investigated in this study. Thus, five limitations apply to the material and the findings of the study: i) the empirical material is restricted to US data, ii) the genre is news texts published on the Internet, iii) the mode is written language in an online setting, iv) the temporal scope applies as above, and v) the results apply to patterns of blend formation from common nouns as they are defined in the present study.

1. Definitions

Many attempts have been made to define lexical blends, but as matters stand only preliminary attempts are at hand [Bauer 2012; Renner et al. 2012; Beliaeva 2014], and Kemmer [2003: 71] argues that “[l]exical blends are so varied in form that no neat taxonomy can do justice to the full range of the phenomenon.” Nonetheless, a very brief but relatively uncontroversial description of a lexical blend can be summed up as follows:

A lexical blend is a lexeme

• formed from two (or rarely several) source words, of which either one or both are truncated in the blended lexical construct. If more than one source word is truncated, the final part of the first source word, and the initial part of the second source word are merged.
• partly motivated by the fusion of structure as well as the fusion of conceptual content of its source words, which sometimes includes shared structure, be it graphemically, phonemically, or both.

There are, however, a number of problems with compressed definitions of this type. First, truncation patterns are reportedly diverse, and there is an intriguing interplay between truncation and the sharing of graphemes and phonemes [Kemmer 2003; Lalić-Krstin & Silaški 2018]. For instance, the structure of a blend such as dracula involves an uncertainty as to the origin of the blend. It is, in principle, not possible to tell which of the source words is responsible for the segment drac-, and the truncation pattern can therefore not be determined, at least not as a result of a structural analysis [Kemmer 2003]. Likewise, the blend robocup is ambiguous as regards the segment -cup, and this blend is made further complex by the fact that the first source word is in itself an amalgamated form derived from robot and cop.

Second, the status of the segments differs as well. Transparency typically varies, and it is common for segments to be reused in forms derived from already existing blends [Kemmer 2003; Lepic 2016; Lalić-Krstin & Silaški 2018]. As indicated above, the recycling of already existing blends, or segments of blends, is a field requiring further research, and the workings of the mechanisms behind these processes are still in need of more detailed analyses.

Third, the meaning potential of a lexical blend is inherently difficult to account for. Although there are observations of iconic qualities in blends such as smog [Ronneberger-Sibold 2006] and priming effects in the process of decoding [Lehrer 1996; Gries 2006], the semantics of lexical blends remains largely unexplored. For instance,
only specific aspects of the source words of *dragula* and *robocup* are intended to fuse so as to result in an intelligible blend. This means that the activation of a certain conceptual subdomain, thus metonymically motivated, is often necessary for successful decoding. The title of the present text obviously draws on this principle; evoking the domain of *gold panning* may be interpreted as an intention to assign a high value to the object of investigation. This is, however, not what it is supposed to convey (although the reader’s interpretation is of course not limited to any particular construal of its meaning). As will become apparent in the following sections, the intended meaning is instead oriented towards the demanding process of retrieval, and the selection of a specific aspect of the source expression is used as an illustration of the metonymic qualities commonly observed in lexical blends.

These three aspects – diversity of truncation patterns, dynamicity of blend segments, and rich and complex meaning potential – are successfully accounted for within a Cognitive Linguistics approach, where the prototype organization and the fuzzy boundaries of linguistic categories are acknowledged and addressed [Taylor 2003; Langacker 2008].

Besides the semantic, or conceptual, gap in blend research, methodological discussions concerned with the collecting of blend data are scarce in previous accounts. There are several possible reasons for this. The presumed marginality of blends in combination with i) their *ephemeral* character [Cannon 1986; Lehrer 1996; Brdar-Szabó & Brdar 2008; Lalić-Krstin & Silaški 2018], ii) their *informality* [Cannon 1986; Fandrych 2008; López Rúa 2012], and iii) their *structural complexity* [Bergström 1906; Bauer 1983; Cannon 1986; Bauer 2012; Renner 2015] means that they are very hard to capture in any ordered and transparent way. In addition, despite the claims that blends are increasing, they are nonetheless infrequent in comparison with other lexical processes. As Arndt-Lappe & Plag [2013] point out, previous accounts of blending have not satisfactorily taken frequency into account, which underscores the need for a more robust empirical background to generalizations concerned with the use and distribution of blends.

It could be argued that there is in fact a vicious circle influencing much of earlier research; blends are portrayed as irregular and unsystematic, which presumably makes them difficult to retrieve systematically. As a consequence, they appear all the more ephemeral, informal, and complex, which further underscores the seemingly unsurmountable objective of systematicity.

Fortunately, there are two distinctive trends in contemporary blend research that promise to push the development in new directions; theoretical and technological progress. The theoretical development of Cognitive Linguistics since the 1980s has opened up the field of linguistics for broader, more dynamic, and more inclusive analyses of language, which explains the nature of lexical blends significantly better than structuralist or generativist views on language [Kemmer 2003; Gries 2006; Fandrych 2008; Silaški & Durović 2013]. Furthermore, the framework of CL in combination with digital corpora, powerful software, and precise psycholinguistic tools offer unprecedented opportunities to explore lexical blending in the 21st century.

2. Data and Method

It follows from the aims of the study that the data are not only accounted for in terms of sources and limitations, but also that the process of retrieving data is a core
methodological aspect in itself. The current section is therefore divided into two subsections aiming to reflect this characteristic of the study. Thus, section 2.1. contains general descriptions of methodological choices and delimitations, while section 2.2. is a procedural account describing in detail the various steps of the investigation.

A comment will also be made here on a few terminological topics. First, descriptions of truncation patterns in blends usually employ the letters AB and CD to represent the (typically two) source words [e.g. Beliaeva 2014]. The initial segment of the SW is indicated with A and C respectively, disregarding the length or other characteristics of the segment. The same rationale is valid for the final SW segment represented by B and D. *Brunch* (breakfast + lunch) is therefore an AD structure, whereas *digicam* (digital + camera) is described as AC. As long as this convention is applied as a schematic structural description of truncation, it serves its purpose well. The only diversion from this coding pattern is the added sign “|” in “A|BC|D”, which helps to signal the uncertain characteristics of blends where no structural material seems to be removed (e.g. *stop|tion* from stop + option).

Second, the topic of this investigation is concerned with *lexical blends*. In analogy with *compounds* and *compounding*, these items are seen as instantiations of the process of *lexical blending*. The term *blend* is sometimes used as an abbreviation of *lexical blend*, while other potential terms, such as *conceptual blend*, *speech error blend*, and *syntactic blend*, are never intended in the abbreviated form.

Finally, section 2.2. contains a certain amount of technical terms needed to account for the procedures of the data collection. These terms are kept at a minimum, and definitions are made as generic and brief as possible, focusing solely on their relevance for the present study.

### 2.1. Methodological choices and limitations

Examples of data collection procedures in previous studies contain descriptions such as “[t]he following pages grew […] out of some chance notations of blends made from time to time by the author” [Pound 1914: Prefatory note] and “[m]any of the examples cited here are from these studies; others are of my own collecting” [Algeo 1977: 47]. Although there are examples of more transparent accounts of the origin of the data, especially in more recent studies, there are few investigations that present data collection procedures allowing in depth analyses of the empirical material in relation to other aspects of language. As a result, discussions on, for instance, i) the semantic relation between source words, ii) the distribution of lexical blend characteristics, iii) conclusions concerning the diachronic development of lexical blending, and iv) genre distribution remain topics that often do not go beyond assumptions in need of more robust empirical grounding in order to be developed further. To some extent, this may explain some of the gaps in blend research that have continued to be unresolved issues for decades. According to Fandrych [2008], the seemingly unsystematic nature of blends has even licensed a neglect of the phenomenon in linguistic investigation, especially in the generative tradition. On the other hand, the difficulties to collect these ephemeral, informal, creative, and complex items in systematic ways are indeed considerable, which is why the emergence of digital media and computer processing have proved to be crucial factors in attempts to establish a better understanding of lexical blending. Last but not least, the CL orientation towards language in use implies
that the focal scope encompasses contextual and general cognitive factors. This means that the requirements on the usage data are considerable, but also that the data are potentially informative as regards the processes involved with lexical blending.

### 2.1.1. Lexical blends and Internet language

Besides the possibility to benefit from the efficiency of digital tools, the choice to investigate data in the domain of Internet language is motivated by, at least, two linguistic characteristics of the domain itself. First, it has been observed that the traditional division between spoken and written mode does not easily apply to Internet language. Crystal [2011: 17] argues that it can be described as “writing which has been pulled some way in the direction of speech”, and in Crystal [2001] it is discussed whether Internet language should in fact be thought of as a mode in itself, separated from spoken and written representations. Second, there are numerous observations of the non-standard (sometimes even called ‘rebellious’) spelling conventions of the Internet [Shaw 2008]. Conventional views on misspelling are thus challenged by (sometimes ideologically motivated) orthographic representations that divert from traditional norms [Sebba 2003].

Both these aspects, the intermediary nature of the modality and the non-standard spelling, dovetail important aspects of lexical blends. As Lalić-Krstin and Silaški [2018] demonstrate, blends sometimes display a discrepancy between their graphemic and phonemic representation in expressions such as Czechout (Czech+ checkout), and Kjellander [2018] shows how the language users may choose to foreground either phonemic or graphemic efficiency in a pair of related lexical blends rapepublican and rape-ublican. These forms are used similarly with similar meanings (derived from rape + republican), but while rapepublican favors graphemic decoding (because of the gemination it causes when pronounced, see Spencer [1996]), rape-ublican appears less transparent graphemically while pronunciation is smoother.

The freedom of the language user seems thus considerable in the formation and use of lexical blends [Bauer 1983; Cannon 1986]. At the same time, it has been claimed that there are constraints that influence where source words are split [Kubozono 1990; Kelly 1998; Gries 2004b] and in which linguistic domains blend tend to occur [Ronneberger-Sibold 2006, 2012]. In addition, the prosody of the source words has been shown to have consequences for the form that a blend takes [Arndt-Lappe & Plag 2013; Beliaeva 2014]. Despite such seemingly restrictive mechanisms, it would still be odd to claim that a certain blend form is incorrect in a traditional sense of the word. Instead, the freedom of the language users typically leaves room for considerable variation, while the observed structural patterns are best described as probabilistic bottom-up tendencies identifiable in their shape [Gries 2012]. The creativity and flexibility of blend formation seem thus to be particularly well aligned with similar observed tendencies in Internet language. It should therefore come as no surprise that Internet language is often referred to as a fertile ground for the formation and use of lexical blends [Lehrer 2007; Cook & Stevenson 2010].

### 2.1.2. Limitations

Previous observations of the distributional characteristics of blends have shown that they appear more markedly in certain domains and genres, e.g. advertising, electronic...
media, and various types of informal contexts [Cannon 1986; Ronneberger-Sibold 2006; Lehrer 2007; Fandrych 2008]. Because of this uneven distribution, the current study assumes that it cannot be taken for granted that observed patterns in lexical blends can be automatically transferred from one domain to another. Instead, it is considered important to delimit the field of investigation in this respect, with the objective to study systematically one relatively delimited language domain, and at a later occasion possibly compare the findings between genres so as to arrive at more wide-ranging conclusions. In addition, the same data collection procedure could be applied comparing different temporal limitations to explore possible diachronic patterns empirically, and the regional scope could be exchanged so as to compare the output from several parts of the English-speaking world.

The chosen genre limitation of the present study is English written news media, which has been claimed to harbor a certain amount of lexical blends [Lehrer 2007], and was therefore considered a good starting point. For reasons presented in the previous section, the empirical material was furthermore limited to online data. It can, however, easily be imagined that these limitations imply a quantitative level of information management that potentially exceeds any reasonable primary estimation of workload. Therefore, it was necessary to work with a systematically organized corpus including inbuilt regional and temporal limitations, and to this purpose the Brigham Young University Corpus News On the Web (NOW) was chosen to constitute the source of the data.

Finally, in order to further limit the output, it was decided to take common nouns as a starting point for the investigation. Apart from general observations of grammatical functions of collected lists of blends and their source words, the interplay between grammatical categories and lexical blends also remains a field in need of further research. Although the present investigation can say little about this aspect within its designated scope, a well delineated dataset in this respect means that it can in principle be compared with future studies aiming at research questions along these lines.

2.2. Data collection procedure

The data source of the study, the NOW corpus, is a digital corpus continuously collecting texts from 20 English speaking regions in the world. The online interface allows various types of restrictions and filters to enable a wide array of linguistic queries related to, for instance, temporal, regional, grammatical, and lexical limitations. Figure 1 exemplifies a simple filtered online query for the word blend used as a verb in the US region.

In the present study, however, the requirement for detailed and multifaceted query restrictions resulted in unstable data outputs and repeated server breakdowns, which eventually led to the decision to download the entire corpus for offline management. This would ensure the stability of the data as well as providing a sufficient level of computer processing power needed to run required queries.
A download was thus performed in April 2018. At this time, the size of the corpus was approximately 6 billion words, and the temporal scope ranged from January 2010 to March 2018, which meant that 99 months of collected online data was included in the empirical material. The corpus came in compressed folders containing text files tagged for region, year, and month (e.g. “15-02-us.txt” for February 2015, US data), which made limitations of these variables easy to manage. Only the data from the United States were retained for the purposes of the current investigation, which resulted in a size reduction to slightly more than one billion words. The NOW data were then imported into SQL-tables in the PostgrSQL management software PGadmin. These operations were necessary as the offline version of the corpus did not come with a user interface such as the one available in the online version (as represented in Figure 1 above).

Another difference between the online and offline data was the data restriction imposed on the downloaded corpus. In short, this meant that 5% of the corpus data were randomly selected and destroyed before download in order to ensure that the material would not infringe on the copyright of original news media publishers. The data destruction is described on the corpus website as a procedure in which “[e]very 200 words, ten words are removed and are replaced with ‘@’”. The corrupt data had therefore to be tracked and removed from the SQL-tables that would contain the corpus. Given the size of the corpus, this was, however, estimated to have only a marginal effect on the query results.

After the necessary arrangements were made in the SQL management environment, a random selection among the 2 000 most frequent common nouns in this part of NOW was performed which resulted in 100 lexical items constituting the base for truncated queries that were intended to be performed at a later stage. The limitation to the 2 000 most frequent nouns was motivated by the great number of items with little or no significance (for the present study) in the so-called long tail of infrequent forms. Examples of such items are presented in Table 1. The column PoS shows the grammatical function (i.e. Part-of-Speech) assigned by the automatic speech tagger CLAWS used by the NOW corpus.
Table 1: Long-tail forms excluded from the data.

<table>
<thead>
<tr>
<th>Wordform</th>
<th>PoS</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guild/AFTRA</td>
<td>nn1</td>
<td>13</td>
</tr>
<tr>
<td>Liiiiittle</td>
<td>nn1</td>
<td>2</td>
</tr>
<tr>
<td>Blues--</td>
<td>nn2</td>
<td>2</td>
</tr>
<tr>
<td>Dramatics</td>
<td>nn1 nn2</td>
<td>2</td>
</tr>
<tr>
<td>WavSTAT</td>
<td>nn1_vv0_np1</td>
<td>1</td>
</tr>
<tr>
<td>p-commerce</td>
<td>nn1_ij</td>
<td>1</td>
</tr>
<tr>
<td>Chih-ping</td>
<td>nn1</td>
<td>1</td>
</tr>
<tr>
<td>EYE/iStockPhoto</td>
<td>nnu</td>
<td>1</td>
</tr>
<tr>
<td>$6,495</td>
<td>Nnu</td>
<td>1</td>
</tr>
<tr>
<td>LARIBA</td>
<td>nn1_np1</td>
<td>1</td>
</tr>
</tbody>
</table>

The randomly selected 100 items were called candidate words (CaW) as they were intended to be the potential lexical sources in the retrieval of lexical blends. The first five CaWs in this list were testimony, track, prosecutor, routine, and citizen. The following step was to truncate the CaWs in order to prepare them for wildcard queries in the corpus. This truncation could not be done in an arbitrary fashion as the output from the CaWs would then depend on a variable beyond systematic control, i.e. comparisons would not be possible within the present dataset, nor with other datasets in potential future investigations. In other words, the truncation should be done according to a model employing the same rationale for all CaWs. The selected model for truncation needed therefore to be dependent on some identifiable and measurable variable in the corpus. To meet this end, the notion of selection point (SP) [Gries 2006] was chosen as it is a systematic identification of a breakpoint based on corpus frequency measures.

In short, the SP of a lexeme is located at the point at which a specific truncation generates a top frequency ranking of the CaW in question. Table 2 illustrates this with the CaW scope*. The point at which scope is truncated to scop* is, thus, where this truncated form renders scope as the most frequent item in the corpus query output (see fourth row in Table 2). The top frequency ranking of scope is, furthermore, taken to be a measure of its prominence in the data. To exemplify the actual output from the SP analysis, the ten truncated forms derived from the five first CaWs were testim*, *imony, trac*, *rack, pros*, *utor, routi*, *tine, citi*, and *zen.

The abbreviation wID in Table 2 is short for word ID, which is a categorization in the corpus roughly comparable to word type, although distinctions between wIDs include grammatical function (PoS), and is therefore more fine-grained than a traditional definition of type. Token frequency, however, is straightforwardly the number of items of
the wIDs in question found in the corpus. The column frequency rank shows the position of the CaW when the corresponding truncated form is employed in a wildcard query*, e.g. if the truncated form sco* is used in a query, the CaW scope is number four in the frequency ordered output list. In addition, it should be noted that even the entire CaW was subject to the truncation script query, which implies that no restriction was set on the possibility for an entire CaW to be the basis for further queries.

Table 2: Calculation of selection point (cf. Gries [2006])

<table>
<thead>
<tr>
<th>Truncation of scope</th>
<th>wID Frequency</th>
<th>Token Frequency</th>
<th>Frequency rank</th>
<th>Selected form for truncated queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>s*</td>
<td>131 145</td>
<td>60 659 869</td>
<td>393</td>
<td></td>
</tr>
<tr>
<td>sc*</td>
<td>10 191</td>
<td>2 789 006</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>sco*</td>
<td>1 167</td>
<td>362 449</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>scop*</td>
<td>55</td>
<td>19 681</td>
<td>1</td>
<td>scop*</td>
</tr>
<tr>
<td>scope*</td>
<td>28</td>
<td>19 445</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Identifying the SP of the 100 CaWs (from both left and right) meant that the total number of truncated forms, henceforth referred to as truncforms, was 200. These were then used in corpus queries producing 200 lists of lexical items potentially containing some proportion of lexical blends. The entire data collection process is illustrated schematically in Figure 2. All steps from download of the database material to the queries based on the truncforms were automatic procedures following the predefined algorithm as described in this section.

Once the lists of lexical items generated from the last step (the truncform queries) were put together in word lists, the remaining analysis (see “Word lists resulting from corpus queries...” in Figure 2) was done manually. In the manual as well as the automated analyses, the specified limitations necessarily applied, which means that while the output was potentially productive, loss of data was inevitable. The strength of the procedure lies, however, in its transparency, which means that it can be replicated relatively easily, but also that contrasting datasets may provide further information in comparative studies.
3. Results

The output lists contained 33,256 lexical items (types) in total, which were analyzed manually. The wildcard query outputs based on all 200 truncforms ranged from 2 to 3,778 wordforms each, but the distribution of these was very uneven, which is illustrated in the mean output number of 166 wordforms, in comparison with the median number of 65. There were, in other words, a number of outliers that diverted significantly from the majority of the queries. For instance, the truncform testim* (from testimony) resulted in only 6 items in its respective wordlist, while dr* (from drive) generated 1,754 items. Notably, the number of blends in these lists were 1 and 6 respectively, which illustrates the unpredictability of the content in the lists.

Figure 3 displays a graph representing the total distribution of the 200 wildcard queries, and as a further illustration the 10 queries (out of 200, i.e. 5%) with the highest number of output constituted more than 44% of the total number of output word forms (14,720 of 33,256). In addition, there was a slight overweight (55%) of wordforms produced by right hand side truncforms, i.e. in which the final, or rightmost, segment of the CaW was removed (hereafter referred to as RT forms, e.g. scop*, as opposed to LT forms from which the initial, or leftmost, segment was removed, e.g. *cope).
The analysis of the output aiming at identifying lexical blends in the data employed the following rationale [Algeo 1977; Gries 2004a; Beliaeva 2014]. Query output items were considered lexical blends if they included an initial truncated segment from SW1 and a final truncated segment from SW2, e.g. brunch (breakfast + lunch). If only one of the SWs, either initial SW1 or final SW2, were truncated, e.g. webinar (web + seminar) or testimoney (testimony + money). If none of the SWs were truncated but instead shared lexical material (overlap), e.g. stoption (stop + option). If the blending of the SWs was non-linear but yet displayed truncation and/or overlap, e.g. Caleavefornia (California + leave). If the orthography of at least one of the SWs was modified and one or all of criteria no. 1-4 still applied.

The total number of lexical blends identified in the corpus output wordlists amounted to 78 items, or approximately 0.23% of the output. Importantly, this proportion should only be seen as a coefficient derived from the observed type frequency. Although its use is limited in this particular investigation, it may be a potentially powerful means to compare future datasets analyzed in a similar way. Furthermore, 46 of the 78 blends (59%) were found in lists generated from RT forms, which is partly explained as a result of the fact that there were already more of these wordforms than the ones generated by left hand side truncated CaWs (i.e. LT forms).

The above presented criteria for deciding what is a lexical blend were used consistently in the entire identification process, but as is shown in previous sections, drawing sharp categorical boundaries must necessarily involve decisions that include, as well as exclude, categorically unclear items [Brdar-Szabó & Brdar 2008]. 62 items in the data
were therefore put in a separate category for future reference, including forms such as petrodollar and dollywood. Although truncated according to a common blend pattern, petrodollar resembles rather a clipped modifier and a head, which seems more reasonable to interpret as an instance of compounding. Dollywood, on the other hand, may evoke the concept of Hollywood, but as the segment -wood is quite frequent in US place names (for instance Greenwood, Westwood, Englewood, Fleetwood, Glenwood, Wynwood, Stanwood) the overlap effect, or wordplay function [Renner 2015], was considered too marginal to include dollywood in the list of blends. Nonetheless, both these examples display the well-known difficulties to classify blends on the basis of their form [López Rúa 2004; Beliaeva 2014; Renner 2015].

Employing such explicit criteria to classify blends may seem counterintuitive from a CL perspective. There are, however, two important points to be made in this respect. First, a distinction is made between membership and representativity in a category [Geeraerts 2006c]. In other words, while a decision needs to be made as to whether a lexical item is considered a lexical blend or not, actual instantiations display blend characteristics to a higher or lower degree. Second, the distribution of blend characteristics is not even. For instance, a semantic analysis of the blend relationship + shit (relationship + shit) has likely a humorous or satirical component, whereas its degree of structural overlap is moderate. In contrast, palimony (pal + alimony) is comparable in terms of degree of structural overlap, but its humorous potential is probably considered rather low. In Cognitive Linguistics, degree of representativity and uneven distribution are typically associated with the idea of family resemblance [Wittgenstein 1968; Rosch 1978; for a specific comment on blends see Brdar-Szabó & Brdar 2008].

Applied to the data of the present study, diverse items such as webinar (web + seminar), crooklyn (crook + Brooklyn), and herstorical (her + historical) are both considered lexical blends although they differ in terms of structure and meaning. For instance, while herstorical elegantly employs a non-morphematic breakpoint to highlight an unexpected aspect of the source word history (from a lexical point of view), webinar resembles a compound web seminar in analogy with blends such as motel without adding any significant level of wordplay and wittiness.

A notable observation in the context of the present study is that more than half of the identified lexical blends (42 out of 78) were not formed from the CaW that had generated them in the corpus queries. For instance, the RT query based on the CaW body (in which the segment bod- was applied) produced the lexical blend bodacious, commonly thought of as bold + audacious, thus following a well attested pattern of semantically related source words such as chill + relax in chillax and fantastic + fabulous in fantabulous [Algeo 1977; Kelly 1998; Gries 2012; Beliaeva 2014; Renner 2015]; i.e. there seemed to be no direct and necessary correspondence between the CaW body and the actual SW of the blend (i.e. bold). This may not strike as particularly surprising given the nature of the data collection procedure, but, at a closer look, these data are potentially informative. Notably, the segment used for each query was selected because it would primarily generate the CaW, at least on the basis of its frequency in the corpus. In the example of bodacious, there are attested sources of an ambivalence as to its first source word, i.e. bold vs. body [Prof. Maarten Lemmens, personal communication], but the frequency measures observed in the current study could explain why this is the case. Contextual analyses of bodacious in the data reveal that the otherwise rather semantically wide sources bold and audacious are narrowed metonymically in the blend so as to associate to aspects concerned with sexual categories. This is reflected in the
observation that the word body is quantitatively prominent, which is accounted for qualitatively as activation of a metonymically restricted conceptualization of body. In addition, the clipped version bod (from body) often used in colloquial American English further stresses the activation potential of the first syllable as body in bodacious [Prof. Raymond W. Gibbs, personal communication].

Figure 4 illustrates the interplay between the lexical items body, bold, and audacious, which make up the conceptual basis for the blend bodacious. As the figure indicates, the conventional use of the blend is constrained by all three lexemes, despite the uncertain role of the lexeme body in the blend’s etymology.

Figure 4: Metonymically derived associations in bodacious

The pattern observed in bodacious is further exemplified in the blend compuware, which is a brand name of a large US based IT company. Figure 5 is a graphical representation displaying the similarity in terms of semantic structure. The relation between the observed SW computer and the CaW company is, however, explicit in this case, which does not influence the structure per se, but instead adds to the information of the characteristics of the blend.
Among the blends with a discrepancy between CaW and SW were also forms such as crooklyn (crook + Brooklyn) RT derived from the CaW crowd. The data did not indicate any association between crook and crowd that would have allowed an interpretation in analogy with bodacious and compuware, which is not surprising given their different phonemic representations (/kɹaʊ-/ vs. /kɹʊ-/). Thus, the phonemic structure did not afford the specific semantic structure expressed in Figures 4 and 5.

An example of an intermediate case in this respect was eco-logical (ecological + logical) generated from the CaW local, which potentially involves a semantic link between environmental concerns and the benefits of the local community in terms of food production, sharing economy solutions, and travels. Although the conceptual aspects of the blend were readily identifiable, the structural properties of eco-logical were considered too generic to support a straightforward semantic explanation of the bodacious and compuware type. The uneven distribution of characteristics noted concerning structural aspects is thus reflected in this property. Furthermore, while the conceptual ambiguity shares the uneven distribution of, for instance, structural overlap, there are no indications in the data of a correlation between these properties in terms of distribution. To illustrate with the examples above, bodacious displays a phonological overlap between the SWs (depending, however, on variation in pronunciation), whereas no overlap is identified in compuware. Eco-logical, on the other hand, which has some degree of conceptual ambiguity in terms of a relation between the CaW local and the blend, displays considerable SW overlap. Crooklyn is also clearly a blend relying on structural overlap, but its qualities as regards conceptual ambiguity between the blend and potentially contributing domains besides the SWs are indeed meagre. Indeed, the conventional understanding of family resemblance is clearly visible in these cases as the characteristics are varying in terms of degree at the same time as their distribution in the individual cases is uneven [Ungerer & Schmid 2006].

Furthermore, approximately 9% of the blends (7/78) displayed a discrepancy between their phonemic and graphemic representation in the sense that the orthography was
necessary to distinguish the blend from one of its source words. Examples in the data were websight (website + sight) and robocup (robocop + cup). There were also examples of striking similarity between the blend and one of the SWs, such as in testimoney (testimony + money). Indeed, the similarity between the syllables /-moʊni/ and /-mʌni/ are so close that it cannot be precluded that the difference in rapid spoken conversation may potentially be confused. Therefore, although there were phonemic cues likely revealing the distinction between testimony and testimoney, the effect of similarity in the blend is hardly a random achievement (cf. Kelly [1998]; Gries [2004a]; Fábregas & Scalise [2012]).

54 The blend, and artist name [López Rúa 2010, 2012], Misstake (Miss + mistake) exemplified both patterns described above in that multiple possible meanings of miss are possible in the meaning construal of the blend, while at the same time the pronunciation alone does not reveal the blend as such. The overt multiple reference creates a richness of semantic potential that is not resolved in the blend, but instead retains ambiguity on several levels; source word origin, semantic potential, and tension between graphemic and phonemic representation. In addition, the powerful dynamics of the initial segment miss foregrounds the analyzability of the final segment -take, which is straightforwardly an English morpheme albeit not typically in the lexeme mistake.

55 As regards truncation patterns among the 78 blends, the quantitative prominence of RT forms in the output, which implies, in practice, that the first SW is truncated, was not reflected in the structure of the blends. Almost half of the blends (46%) were of the ABD type (e.g. webinar from web + seminar), i.e. forms in which SW1 was retained in full. The other half consisted of ACD and AD blends distributed relatively evenly, with only three examples of the A|BC|D type (palimony from pal + alimony, popera from pop + opera, and eco-logical from ecological and logical), in which truncation does not seem to follow the typical pattern. These proportions are represented in Figure 6. In other words, there was a marked tendency in the data towards preservation of the first source word, which is especially interesting since there was an output tendency going in the opposite direction.

56 It is also notable that the AD structure, which is traditionally seen as a prototypical truncation pattern in blends [Plag 2003; Gries 2004a; Brdar-Szabó & Brdar 2008] constituted only a fourth of the blends in the data. The centrality of an AD pattern in a prototype categorization account of blends need not necessarily be questioned for this reason, even if the findings of this investigation would prove robust in further studies. From a quantitative perspective, on the other hand, it seems problematic in relation to the observation in Beliaeva [2014: 33ff.] that “it is widely accepted […] that the majority of blends combine the initial part of the first source word (W1) with the final part of the second source word (W2); in terms of the formula […] AB + CD = AD”.

Lexis, 14 | 2019
There were also examples of so-called *intercalative* blends in the data. These are blended forms “in which the two source words involved in the blend are so tightly integrated in the blended word that the sounds of one source lexeme are interspersed between the sounds of the other” [Kemmer 2003: 72]. The ambiguity of such blends is recognized in Kemmer [2003], but in the context of the present study this characteristic is of particular interest. Examples from the data were *dragula* (drag + dracula), *construction* (construction + action), *testimony* (testimony + money), and *ecosystem* (ego + ecosystem). The structure of the previously mentioned blends *misstake*, *websight*, and *robocup* was also of the intercalative type, although these items were more complex, and contained aspects that add to the intercalativity.

It was further noted that metonymy and metaphoricity were recurrent semantic characteristics in the formation of the blends in the data. In the conceptual domain *web* (in the metaphorically derived sense related to the Internet domain), the blend *websight* highlights *seeing* and *being seen* as an aspect of the role of a *website*, i.e. websites usually aim at attracting Internet users, which is typically profiled in visual categories. Importantly, this observation requires a familiarity with the Internet domain, especially since there is a literal sense of *web* that could, at least under certain circumstances, be a potential competitor SW. It follows from this that observing the figurative qualities depends (in this case perhaps apparently) to some extent on encyclopaedic knowledge in the construal of the blend’s semantic content. The figurativity of such blends signals thus the dependence on other types of knowledge than what has traditionally been thought of as *linguistic* knowledge.

Similarly, ethical considerations of the conceptual category *testimony* are selected as *money* is foregrounded in *testimoney* and gender inequalities in the designation of *historical* accounts are profiled as *her* is introduced in *herstorical*. The latter example
highlights, furthermore, the influence of creativity in relation to conventional morphological patterns. For reasons of space, the figurativity in lexical blending is not operationalized further in this study. Suffice it to say presently that figurativity is a neglected and potentially important semantic aspect in need of further investigation.

4. Discussion

The number of lexical blends that were identified in the 200 query output lists may appear rather low, especially given the numerous reports on the increasing popularity of blends [Bryant 1974; Kemmer 2003; Lehrer 2007; for a critical comment on this topic see also Lepic 2016]. Still, few studies have looked into this issue systematically enough to suggest clear definitions what is meant by increasing, let alone to present any figures behind such a presumably measurable change. Based on the findings of the present study only, it is therefore not possible, at this point, to say anything conclusive about the role of blends in natural language use, especially from a diachronic viewpoint. An investigation of a comparable dataset from the beginning of the 20th century (whether that would be obtainable at all) could possibly answer such questions, but as matters stand, this remains largely a field for speculation. Moreover, the limitations in terms of region and genre (especially the latter) have likely implications for the number of blends in the data of this investigation, which is in itself a call for further studies and replication of the current methodology on other datasets.

The problematic issue of defining what is a lexical blend (and not) further complicates the quantitative analyses [Bauer 2012]. One perspective that may contribute to developing the notion of category delineation is to take a prototype theory perspective [López Rúa 2004; Brdar-Szabó & Brdar 2008; Bauer 2012; Renner et al. 2012]. It follows from such a viewpoint that holistic analyses of the objects under investigation are performed, and that attributes, or characteristics, are studied in relation to their distribution among the data. The conventional way in Cognitive Linguistics is to interpret such relations in terms of family resemblance [Rosch 1978; Geeraerts 2006b; in the context of lexical blending see Brdar-Szabó & Brdar 2008], which enables comparisons with naturally occurring categories such as bird and dog. Methodological systematicity requires nonetheless that

a strict distinction has to be made between degree of membership and degree of representativity. Membership in the category bird is discrete; something is or is not a bird. But some birds may be birdier than others: the swallow does remain a more typical bird than the ostrich. [Geeraerts 2006b: 150]

For the purposes of methodological design, definitions need therefore to be based on membership categorization, while analyses and discussions need to consider the degree of representativity as a central notion. In other words, no single study is likely to resolve the riddles of lexical blending. Instead, patterns emerging from a multitude of studies relying on systematic and replicable methodologies have better chances to increase the level of knowledge.

The analysis of the blend list revealed a few notable properties of the data as regards truncation. First, the majority of the blends (72% excluding A|BC|D blends) were formations in which only one of the source words was truncated. Second, although the initial queries resulted in significantly more word forms based on a right truncation (RT) of the CawS, almost half of the blends were formed with the first source word
retained in full. Third, the widely spread idea that the AD structure should be the most prevalent pattern in lexical blending [Beliaeva 2014] was not supported by the data in this study, which is perhaps the most striking, and equally most problematic, observation of the truncation patterns. Furthermore, as news media (which is the origin of the data in the present study) have been claimed to be one of the central domains in the development of lexical blends [Bryant 1974; Lehrer 2007; Fandrych 2008] the expectation would have been that AD blends would have constituted a prominent part of the data. This was apparently not the case, and the observed discrepancy needs to be followed up by further investigations. Already at this point, however, it raises questions as to the nature of the data in previous studies. For instance, if no systematic rationale is employed in the data collection, it may be tempting to focus on powerful and witty blends, as these may seem to display a higher density of blend characteristics, and would therefore possibly be considered more representative. The blend *misstake* in the present study exemplifies such a formation employing several strategies, especially in terms of semantic content, and it may thus, on good grounds, be claimed to be highly significant as to qualitative description. This does not necessarily mean that it is representative of lexical blending as a process. Datasets that do not take this issue into account run therefore the risk to present lexical blending in a biased way. Systematic data collection seems thus ever so important in further accounts of this word formation process. The validity and the relevance of the findings are of course central notions in this respect, but perhaps equally important is the opportunity to ask questions about blends that have previously been difficult to address.

The observations of ambiguity in the blend data are in many senses parallel with the reasoning on wordplay in Renner [2015]; the emergence of various types of the phenomenon is more than haphazard consequences of certain formal structures. Although unevenly distributed in the data, both wordplay and ambiguity seem to be consistently, and functionally, employed as motivating attributes creating affordances for language users to form and use lexical blends. A preliminary division of the types of ambiguity found in the data in the present study can be described as i) source word ambiguity, ii) conceptual ambiguity, iii) mode ambiguity, and iv) truncation ambiguity. These attributes have the potential to co-occur freely in any instantiation, and the boundary between them is typically fuzzy.

First, source word ambiguity (e.g. *dra-* in *dragula*) have, for instance, been observed by Kemmer [2003] and Bauer [2012], but the functional properties of this characteristic have not yet been related to motivations for lexical blending. The ambiguous structure has rather been discussed as an analytical problem, which it is indeed from a structural perspective. The distinction between intercalative and non-intercalative structure is in fact questioned in Gries [2004b], where the apparent inconsistency is identified as the possibility for either analysis in Kemmer's [2003] data. In contrast, it is argued in the present study that Kemmer [2003] has actually spotted an important motivation, or a cognitive constraint, for the formation and use of blends. Put differently, what has been perceived as a structural anomaly, or even an analytical flaw, is instead claimed here to be an important hint to the nature of lexical blending.

Second, the possibility for indirect sources to influence the meaning construal in blends such as *bodacious* and *compuware* constitutes another type of ambiguity, which profiles the activation of a third domain of meaning in blends that appear to have only
two sources. In the present study this characteristic is referred to as conceptual ambiguity. This type of ambiguity was relatively rare in the blend data, and further studies are needed to account for the distribution in detail. This strategy displays a striking similarity with figuratively motivated meaning construals as regards semantic structure; conceptual content is captured from an indirect domain, be it metonymically, metaphorically, polysemously, or even homonymously associated to the primary, or overt, source domains (SW1 and SW2). Given the common pattern of metonymically motivated combinations of concepts, the mechanism of conceptual ambiguity is not a surprising finding.

Third, the importance of phonology for lexical blending has sometimes been highlighted in previous research [Cannon 1986; Kubozono 1990; Lehrer 1996; Kemmer 2003]. Interestingly, the findings of the current investigation also underscore the orthography as a significant aspect of blend formation. While phonology is often related to structural concerns such as breakpoints, sound similarity, and prosody [Cannon 1986; Kelly 1998; Gries 2004a, 2006; Arndt-Lappe & Plag 2013], the functional use of spelling in the creation of blends seems to draw rather on semantic aspects, at least as regards the blends in the data of the present study. Moreover, the result of the discrepancy between orthography and phonology is understood as a dynamic interplay between these modes of language. The simultaneous activation of website and sight in the blend websight is thus the result of the synchronously produced ambiguity of the graphemic and phonemic representation, and the term chosen for this mechanism is subsequently mode ambiguity.

Finally, the blends palimony, eco-logical and popera illustrate a commonly observed pattern in lexical blending [Algeo 1977; Beliaeva 2014]. A conventional way to describe such items is to argue that the source words are retained in full in the blend (i.e. no truncation has occurred), and as this central characteristic is absent it may be tempting to exclude these items because they lack a central property (see for instance Ralli & Xydopoulos [2012]). This is questionable for at least two reasons. First, there is a growing understanding of lexical blends in terms of prototypically organized categories [López Rúa 2004; Brdar-Szabó & Brdar 2008; Bauer 2012; Renner et al. 2012; Renner 2015], which implies that identification and analyses relying on formal semantics listing of necessary and sufficient conditions are inherently misleading. It has even been argued that such approaches are (at least partly) responsible for the 20th century failures to account successfully for the complexity of lexical blending [Fandrych 2008]. Instead, holistic analyses of objects, including asymmetric attribute distribution, have proven much more apt to explain the complex nature of blends [Brdar-Szabó & Brdar 2008; Bauer 2012; Renner et al. 2012]. Thus, although truncation is seen as an important attribute of lexical blends, it does not have the status of a necessary condition, simply because blends cannot be categorized satisfactorily along such lines.

Second, the idea that both source words are retained in full in blends such as popera overlooks the structural and semantic realizations of ambiguity as a functional cognitive constraint. The analyses of the present study reveal several types of ambiguity as important motivations for lexical blending, and popera exemplifies one such ambiguity, albeit manifested in its structure rather than its semantics. It is therefore more reasonable to interpret its truncation as a structural ambiguity, or truncation ambiguity, produced as a result of the dynamicity between pop and opera as they are simultaneously represented in the blend. Put simply, while pop is profiled,
opera is backgrounded, and vice versa [Ungerer & Schmid 2006; Langacker 2008]. There is thus an apparent analogy with the well-established notion of binocular rivalry as a visual version of multistable perception [Levelt 1968; Rubin 2003] (Figure 7). Neither focus of attention allows a simultaneous attention to the contrasting object, which is realized in popera as a dynamically construed truncation, i.e. either of the SWs pop and opera are truncated depending on which of them is foregrounded.

Figure 7: The face-vase illustration of binocular rivalry

In all, one of the keys to separating lexical blending from other word formation processes may in fact be the functional use of ambiguity and uncertainty. From the perspective of similarity between SWs and between SWs and the blend, Gries [2012] indicates this by pointing out that if too much of the SWs are retained in blends, they are not likely to be successful simply

because they are not fun anymore: while both source words are perfectly recognizable, the blend is not too similar to either source word anymore and the punning/playful character of such blends is largely lost. [Gries 2012: 159]

This reasoning on similarity is easily transferred to ambiguity as it is identified in the present study. As mentioned in the introduction, punning in this context is understood broadly as a concept drawing the attention to the idea of pleasure in the shape of intellectual challenges. The explanation for the impact of ambiguity, and indeed a motivation for the existence of lexical blends, is thus found in parallel with other intellectual challenges such as jigsaw puzzles, crosswords, and riddles.

The four types of ambiguity described in this section combine to a powerful cognitive constraint influencing, or even driving, the development and use of lexical blends; namely multistable meaning construal as an instantiation of the well attested cognitive
mechanism of multistable perception [Leopold & Logothetis 1999; Rubin 2003; Kornmeier et al. 2009]. Moreover, the underlying motivation for this constraint in blending is identified as a manifestation of a drive to seek intellectual challenge.

This explanatory frame encapsulates several findings of the blends in the data of the present study, but it is also in line with previous research, and moreover, it dovetails the Cognitive Linguistics claim that no sharp boundaries can be identified between language and other cognitive abilities and processes. Although lexical blends may be claimed to be marginal in comparison with word formation processes such as derivation and compounding, and more so yet in relation to lexical retrieval and syntactic processes in general, the findings in lexical blend research may have far-reaching implications for more general linguistic assumptions [Kelly 1998; Dressler 2000]. Also, outbursts of lexical creativity seen in controversial political circumstances [Lalić-Krstin & Silaški 2018] signal that the importance of lexical blends go far beyond the indications given by their low frequency numbers. Therefore, it seems even more pressing to take further steps towards a better understanding of these lexical items.

5. Conclusion

The two aims of the study – on the one hand an approach to data collection and methodology, and, on the other hand, an investigation of lexical ambiguity in the formation and use of blends – are in several ways indications for further research. The data collection methodology has provided important insights into the possibilities as well as difficulties in systematic collection of lexical blends. Some of the quantitatively oriented findings raise central questions about traditional conceptualizations of the nature of blends, while other observations need the context of further studies to be assessed as to their explanatory value. It seems clear, however, that Bauer [2012: 21] can be paraphrased successfully in the context of the present study by exchanging ill-defined with unsystematically collected in the phrase “[w]e need more than new experiments on an ill-defined set of words”. Bauer’s [2012] original intention in this quote contains still an unresolved issue, but systematicity is at the heart of further development not only as to general epistemological concerns, but also for the development of definitions and classificatory devices. As the present study shows, there is a potential in enriching the qualitative analysis with quantitative data in that previous conceptions of irregularity and complexity may be less dramatic in taxonomic matters than has been traditionally presumed. Also, it is demonstrated in this study how semantic explanations, such as the suggestion to understand ambiguity as an intentional process along the lines of multistable perception, have consequences for structural analyses as well.

The recurring claim that broad approaches are needed to explain lexical blending [Kelly 1998; Fandrych 2008; Renner et al. 2012] is further illustrated in this study. Categorization, classification and structural properties have been addressed in a multitude of accounts, but the semantic characteristics of blends constitute still a challenge to linguistic investigation. Methodological creativity is likely central in this respect. Kelly [1998] acknowledged this need two decades ago, and the current investigation hopefully adds another piece to the puzzle of lexical blending.
BIBLIOGRAPHY


BRYAN Margaret M., 1974, “Blends are increasing”, American Speech, 49, 163.


NOTES

1. This does not exclude the possibility for blends resulting from a speech error to be adopted in language use as instances of intentionally motivated lexical items. The term word formation blend merely indicates the use of an item regardless of its etymological origin.
3. Structured Query Language (SQL) is a programming language designed to manage relational databases.
7. Table 2 in the present text is a variation on Gries’ [2006: 543] table 2, which uses the example word agitation.
8. The symbol ‘*’ is used here as it is the conventional way to signal wildcard queries. In SQL language, however, the wildcard query symbol is ‘%’.
11. The alternative expression sandwich word to indicate this structure is mentioned in Algeo [1977].

ABSTRACTS

The multifaceted and intermediary nature of lexical blending has been discussed from various theoretical perspectives in the last decades [Kubozono 1990; Kelly 1998; Dressler 2000; Kemmer 2003; Gries 2006; Fandrych 2008]. Corpus linguistic studies have contributed to a growing body of empirical data demonstrating significant patterns of blend formation [e.g. Gries 2004b, 2006; Beliaeva 2014]. These more recent findings are particularly important as they illustrate that lexical blending is not as irregular and unsystematic as it has often been assumed in the past [Lehrer 1996; Kelly 1998; Kemmer 2003; López Rúa 2004; Beliaeva 2014]. Because blends are often short-lived [Cannon 1986; Lehrer 1996] and informal [Bauer 1983; López Rúa 2010; Bauer 2012], making predictions about their nature is a complex task. Systematic analysis involves therefore a number of challenges. One such challenge is how to detect and collect these unspecified and untagged lexical items in a corpus. It is a well-known fact that conventional dictionaries are poor sources [Cannon 1986], and alternative methodologies are therefore required. Another challenge is how to ensure that the collected data is representative of all lexical blends within a selected set of limitations. Besides addressing these challenges, semantic aspects of ambiguity were investigated from a Cognitive Linguistics (CL) perspective [Geeraerts 2006a]. The patterns of ambiguity in the data are explained as instantiations of multistable perception, which is understood as a phenomenon in which “our perceptual system fails to produce a stable unambiguous percept” [Kornmeier et al. 2009: 138]. From a linguistic perspective, this means that
aspects such as phonology, orthography, and semantics constitute variables holding a potential for functionally employed lexical ambiguity [Renner 2015]. At various stages, the study employed both automated processes and manual analyses, which means that workload limitations applied to the scope of the data. A central aim of the study was, however, to develop systematic approaches to the collection of lexical blends, precisely because it is considered necessary in the development of blend research in general.


INDEX

**Keywords:** lexical blending, semantics, corpus linguistics, multistable perception, methodological development

**Mots-clés:** amalgamation lexicale, sémantique, linguistique de corpus, perception multistable, développement méthodologique

AUTHOR

DANIEL KJELLANDER

Umeå University, Sweden
daniel.kjellander@umu.se