Pierrehumbert, Janet B. (in press) Burstiness of Verbs and Derived Nouns. In Diana Santos, Krister Linden, and Wanjiju Ng'ang'a (Eds.) Shall we Play the Festschrift Game?: Essays on the Occasion of Lauri Carlson's 60th Birthday. Springer Verlag. (to appear October 2012).

Burstiness of Verbs and Derived Nouns

Janet B. Pierrehumbert

1 Introduction

The frequencies of words vary with the discourse context. People use different words depending on what they are talking about, and what they understand to be their common ground with their interlocutor (Tanenhaus and Brown-Schmidt 2008). Once a topic of discussion is initiated, it is recursively elaborated until the next topic shift (Kintsch 1974, Sharkey and Mitchell 1985). These facts provide the statistical foundation for modern document indexing and retrieval. When people search documents, they have the goal of finding information about some topic, and words that are statistically concentrated in particular documents, in contrast to others, provide good characterizations of the document topics (Bookstein and Swanson 1974, Church and Gale 1995, Katz 1996). If a word characterizes the topic of document, the likelihood that it will occur a second time in the same document is far higher than its average frequency would predict, because people reuse the word as the discussion is elaborated (Sarkar et al. 2005, Church 2000). In the statistical natural language processing literature, the term *burstiness* is used to designate the tendency of topical words to occur repeatedly in bursts, separated by lulls in which they do not occur because different topics are under discussion.

Good keywords for document retrieval are very bursty. But essentially all words are at least somewhat bursty, in that their temporal distributions display significantly more clumping than would be predicted under a baseline model in which words are strung together at random (Altmann et al. 2009). Even common words that everyone knows, such as *she* or *yesterday* are more relevant and useful for some topics of discussion than for others. This essay inquires into the underlying mechanisms by which different burstiness values come about. In particular, I will be interested in the the relationship between the meanings of words, and the way that words are used in

Janet B. Pierrehumbert
Department of Linguistics, and Northwestern Institute on Complex Systems,
Northwestern University, Evanston, IL, USA, e-mail: jbp@babel.ling.northwestern.edu

discourse, which is the central focus of Lauri Carlson's Ph.D. dissertation and 1983 book (Carlson 1983). I will explore this relationship by comparing derived abstract nouns, such as *discussion*, both to the verbs they are derived from (e.g. *discuss*) and to nonderived frequency-matched simple nouns, such as *pool* and *child*. The derived nouns inherit semantic structure from their base verbs. However, they function in discourse like other common nouns.

2 Background

Quantifying the burstiness of words has been the subject of a substantial research literature, because the phenomenon in itself challenges the assumptions of the conventional statistical treatment of words. To estimate the frequency of a word, we normally assume that the frequency is a permanent (e.g. stationary) property of the word, and that a text sample is a frequency-weighted random sample of the words in the lexicon. The statistical estimate of the word frequency is expected to converge to the true word frequency as bigger and bigger text samples are taken. However, the rate of convergence is poor if the word frequency fluctuates. The larger the scale of the fluctuations, as measured in words of text, the worse the convergence to the true frequency becomes. The common theme in proposals for quantifying burstiness is the addition of one or more parameters to the statistical information about each word. The purpose of the additional parameters is to capture local fluctuations in the word frequencies.

The most common approach takes as its point of departure a division of the text corpus into documents, and seeks to characterize the *counting distributions* of words (the distribution of word counts with respect to documents). This approach is the natural one if the division into documents is self-evident, for example if the dataset is a collection of news stories, research articles, or blog posts. It has been successfully applied not only in document retrieval (Church and Gale 1995, Katz 1996) but also in predicting the dynamics of words over time (Altmann, Pierrehumbert, and Motter 2011) and in psycholingistic studies of word processing (Heller et al. 2010, Heller and Pierrehumbert 2011). However, the approach casts away all information about the way the word is distributed within single documents, and the results can be very sensitive to the way that the dataset has been partitioned, in cases where the division into documents is not self-evident. The present study builds on the earlier study by Altmann, Pierrehumbert, and Motter (2009) that adopts a different viewpoint by exploring the distributions of word recurrence times (the times between one use of the word and the next) in a single long stream of text.

The stream of text for Altmann et al. (2009) was obtained by downloading the archive of the Usenet discussion group talk.origins from Google Groups. This dataset contains approximately $N \approx 2 \times 10^8$ words produced by fifty thousand users who debated evolution and creationism over the time period from September 1986 to March 2008. We take it as a model system for the flow of informal dialogue in a social community. Although it is admittedly topically restricted, the same might

be said for discourse within any social group, as social groups are often brought together by shared interests and experiences. The dataset consists of threads that in turn consist of posts, where each post is an individual communication by a user. The lengths of their posts range from a single word to over 3,000 words. The text was collated by threads according to the time stamp in the first post of the thread. Unequivocal examples of spam and threads in languages other than English were removed, as well as repetitions of parts of previous posts that were marked as such with >, |, etc. The recurrence time distributions of the 2,128 words that occur at least 10,000 times were modeled; it was shown that the single free parameter of the Weibull (stretched exponential) distribution effectively captures the degree to which the behavior of any particular word deviates from the exponential distribution that would be expected if words were strung together randomly. The model is prefigured by Sarkar et al. (2005), but differs in using only one free parameter to achieve an extremely accurate fit (median $R^2 = 0.993$) for recurrence time distributions. It has been independently validated through a comparison to empirical bootstrapping techniques in the context of setting significance levels of keywords for the British National Corpus and the San Francisco Call NewspaperCorpus (Lijffijt et al. 2011).

What determines the burstiness of any given word? Is the value an idiosyncratic property of the word? Is it predictable from intrinsic properties of the word? Or does it arise indirectly from the interaction of the word with the discourse structure? Altmann et al. (2009) focus on an intrinsic property of the lexical semantics, *logicality*, and on a conjecture by von Fintel that *logicality* is correlated with *permutability*. In von Fintel's words, "logical meanings are invariant under permutations of the universe of discourse [...] The intuition is that logicality means being insensitive to specific facts about the world." (von Fintel 1995). He continues to suggest that high *semantic type*, in the sense of Montague (1973) or Partee (1992) is associated with high logicality, and greater permutability. In order to elucidate this tripartite relationship, I first review the theory of semantic types, and then discuss the operationalization of the concept of permutability.

Formal semantics undertakes to provide a compositional treatment of the semantics of sentences, in which the truth conditions for any sentence can be predicted from the semantic components contributed by its parts. In furtherance of this goal, the formal representation of any word includes the domain and image of the mapping that is implicitly associated with it by virtue of its meaning and the constructions in which it appears. For example, proper names such as *Darwin* are treated as entities (in formal notation, they have type $\langle e \rangle$). At first blush, one might imagine that common nouns such as *monkey* are also type $\langle e \rangle$. Compare:

(1) John likes Sue. John likes that monkey.

This impression is deceptive, however, because *monkey* actually refers to the set of monkeys, that is, the set of things that have the properties characteristic of monkeys. In example (1), the demonstrative *that* functions as an operator to single out an entity from this set. From a semantic point of view, common nouns are thus the same as extensional adjectives, and this fact goes towards explaining why so many words can be used in either syntactic role:

(2) Her dress was blue. Blue looks good on her.

Since common nouns such as *monkey* and *blue* correspond to sets, which are equivalent to properties, they can take entities as arguments and map them to truth values t, as in the sentence *Lake Tahoe is blue*. The sentence is true if Lake Tahoe is a member of the set denoted by *blue*, and false otherwise. In formal notation, this is shown as $\langle e,t \rangle$. Essentially quantificational nouns, such as *everyone*, are characteristic functions of sets of properties of entities (van Benthem 1989, Partee 1992).

A given word can have different types in different contexts, because languages have productive processes of type shifting. *Disney* is originally a proper name for a person, of type $\langle e \rangle$, and retains this type as a proper name for the corporation that created Mickey Mouse and Disney World. However, it can be readily understood as a modifier or predicate in the following attested example:

(3) The reason we don't want to Disney is that we do everything Disney ...
(mousepad.mouseplanet.com 11-29-2006)

The listener readily reinterprets the entity *Disney* as the activities and properties that are characteristically associated with that entity.

From the point of view of morphological theory, examples like (3) can be characterized as examples of conversion, or zero-derivation. A proper noun is converted into a predicate without the addition of any phonological material. Its function in the sentence is like that of other predicates, with the result that the apparatus of Montague semantics can integrate the formal representation of *Disney* together with those for the other words in the sentence into a semantic representation of type $\langle t \rangle$ (e.g. the whole sentence is either true or false). This line of analysis is very plausible if part-of-speech conversion is productive and resembles other processes that linguists attribute to the syntax. In languages with rich morphology, however, it can be difficult to draw the line between the compositional syntactic and semantic structures that provide the foundation for Montague semantics, and morphological principles operating within the lexicon. Even for English – whose impoverished morphology has allowed computational linguists to go far with a naive orthography-based concept of a word – psycholinguistic evidence now suggests that many or most morphologically complex words have their own entries in the mental lexicon (cf. Baayen et al. 2007) and that there is no clear dividing line between compositional and noncompositional complex words (Hay 2003). While Montague semantics took words as given, and looked to understand how words combine in larger units, such findings raise questions about how semantic properties are combined and inherited within words.

A central assumption in the cognitive theory of the lexicon is that words tend to have a basic or unified representation (Bybee 2001, Blevins and Wedel 2009). Thus, we can ask what a good starting assumption for the basic semantic type of a word is. Montague took the lexical type of a word to be the highest type in which it occurs, leading to the conclusion that all nouns are generalized quantifiers, just like the term *everyone* (Montague 1973). However many recent researchers, including Partee (1992), take the lowest type as basic. Since type raising is far more productive

than type lowering, classifying words by their minimal type leads to sharper distinctions amongst the various sets of words. Altmann et al. (2009) follow this scholarly trend. In their data analysis, each word is coded by the lowest type in which it commonly occurs. (Exceptional uses, such as *associated* in the proper name *Associated Press* are disregarded in the interests of statistical clarity.) The end result is a ladder of abstraction, along which words are positioned according to the types of logical relations that they manipulate. In the interests of coding reliability, the full type hierarchy is collapsed to four broad classes, as shown by the examples in Table 1.

Table 1 The full type hierarchy collapsed to four broad classes

Class	Name	Examples of words
1	Entities	Africa, Bible, Darwin
2	Predicates and Relations	blue, die, in, religion
3	Modifiers and Operators	believe, everyone, forty
4	Higher Level Operators	hence, let, supposedly, the

Note that the semantic classes are only partially correlated with syntactic parts of speech. Words such as *blue* are treated as predicates or relations no matter whether they appear as nouns or adjectives in a sentence. Any word with an intrinsically quantificational or relational meaning is coded as having a higher type than other words of the same part of speech that lack this meaning component. Notably, scalar adjectives such as *huge*, *former*, *legal* are all coded as Class 3 and not Class 2. The overall coding approach was adopted both for practical and for theoretical reasons. On the practical side, our corpus is much too large to be hand-coded, but automatic part of speech taggers trained on formal prose become less reliable when faced with the short phrases and out-of-vocabulary words of colloquial language. On the theoretical side, the extremely free part of speech conversion in colloquial English can present real challenges to syntactic theory. Further, although correlations of burstiness with part of speech have been reported (Church and Gale 1995, Montemurro and Zanette 2002), no one has put forward a precise proposal about how these correlations might arise. Exploring a semantic point of view is attractive because it offers leverage on the underlying mechanisms for burstiness patterns.

The leverage occurs through the concept of *permutability*. When von Fintel refers to "permutations in the universe of discourse", he appears to have in mind the different ways that a word could be used if one discourse context is substituted for another. A word of high type and high logicality, such as *forty*, has potential relevance to the discourse no matter whether we are discussing onions or some other topic, such as books or houses. This is a paradigmatic view on permutability, in that it deals with the structure of available alternatives in discourse. It can be reinterpreted syntagmatically (e.g. in relation to the sequential structure of discourse) by bringing to bear two assumptions: First, a topic of discussion can be characterized as a probability distribution over sets of words (Blei et al. 2003) and second, as human discourse unfolds in time, it randomly traverses the space of potential topics. Together, these

assumptions imply that randomly permuting all the words in a very large text sample is equivalent to randomly reassigning each word from the discourse contexts in which it occurred to other actual or potential discourse contexts. If a given word was very bound to particular contexts, this reassignment would greatly affect its statistical signature. However, there would be little effect on the statistical signature if the word were not bound to particular contexts.

Randomly permuting all the words in a text corresponds to the so-called Bag of Words model in the statistical natural language processing literature (Nigam et al. 2000). This model can be conceptualized by thinking of each word type as a ball with the word written on it. The lexicon is a large bag, in which the number of balls with a particular label corresponds to the frequency of the word. A sequence of words that results from drawing one word after another out of the bag (with replacement) corresponds to a random reordering of a text from which the lexicon was derived. Equally, it exemplifies what the text would look like in the absence of further factors structuring the discourse. This model is trivially false at syntactic time scales, but provides an important comparison at longer, discourse-level time scales.

If words of high logicality are highly permutable, then their temporal statistics should be relatively unaffected by randomizing all the words over a wide-ranging set of discourse topics. If words of low logicality are highly variable under permutation, their temporal statistics should be greatly affected by this operation. Or, to put it differently, we can look for evidence of logicality in the extent to which a word's actual distribution deviates from the hypothetical distribution it would have in the Bag of Words baseline model. In the next section, I will describe a mathematical apparatus for quantifying these deviations.

3 Quantifying Burstiness

Here, I summarize the method used in Altmann et al. (2009) to quantify burstiness in word recurrence distributions while controlling for word frequency. The recurrence time $\tau_j^w = i_{j+1}^w - i_j^w$ is defined by one plus the number of words between two successive uses i_j^w and i_{j+1}^w of word w, where i is an index running from 1 up to the length of corpus. Under the Bag of Words model, the set of observations of any given word is generated by a Poisson process. The average recurrence time $\langle\tau\rangle$ is the reciprocal of the word frequency ν (eg. $\langle\tau\rangle = 1/\nu$). The recurrence time distribution is predicted to be exponential:

$$f_P(\tau) = \mu e^{-\mu\tau}, \tag{1}$$

where μ is a parameter of the distribution that is equal to the word frequency: $\mu = \nu = 1/\langle\tau\rangle$.

It is important to note that this baseline does not correspond to an even distribution of the word throughout the text. In throwing dice, a number may happen to

come up several times in row or happen not to come up over many trials. In the same way, when a given word is randomly selected under the Bag of Words model, it may have a run of good luck in which it appears repeatedly, or a run of bad luck in which it is rarely selected. The ribbon plot in Fig. 1 shows how a hypothetical word is distributed over time in a sample outcome of the Bag of Words model.

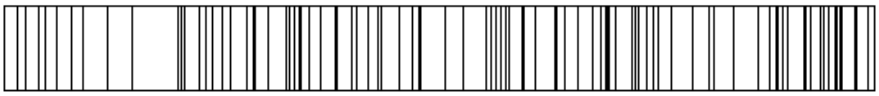

Fig. 1 Occurrences over time of a word with $\langle\tau\rangle = 820$, in a sample outcome of the Bag of Words baseline model. The outcome is shown for text of length $100\langle\tau\rangle$ = 82,000 words. The black lines (whose thickness is exagerated for visual clarity) indicate a time point at which the word occurs. Reproduced from Altmann et al. (2009)

We are interested in deviations from the kind of behavior shown Fig. 1. The ribbon plot in Fig. 2 shows the actual outcome for a fairly bursty word in this dataset. This is the word *theory*, which appears very frequently in talk.origins in connection with arguments about different theories.

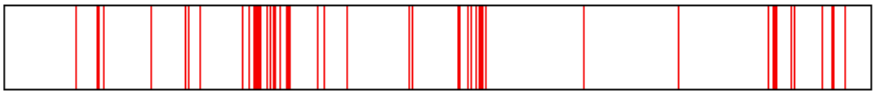

Fig. 2 Actual occurrences of *theory*, a word with with $\langle\tau\rangle = 820$, in a 82,000-word sample drawn from talk.origins. Reproduced from Altmann et al. (2009)

This word is more concentrated at some times than the baseline model in general predicts. These concentrations are balanced by long lulls in use of the word. Next, in Fig. 3 we look at a word of the same frequency and a higher semantic type, namely the word *also*.

Fig. 3 Actual occurrences of *also*, a word with with $\langle\tau\rangle = 820$, in a 82,000-word sample drawn from talk.origins. Reproduced from Altmann et al. (2009)

Clearly, the word *also* is closer to being exponentially distributed than *theory* is. However, there is still a noticable difference between Fig. 3 and Fig. 1, which indicate that *also* is somewhat bursty.

The central finding of Altmann et al. (2009) is that the the distribution of each of these words – indeed of each of the 2,128 words that occurred at least 10,000

times – can be extremely well captured by fitting the single free parameter β of the stretched exponential distribution

$$f_\beta(\tau) = a\beta\tau^{\beta-1}e^{-a\tau^\beta}, \tag{2}$$

where $a = a_\beta = [\nu\,\Gamma(\frac{\beta+1}{\beta})]^\beta$ is obtained by imposing $\langle\tau\rangle = 1/\nu$, Γ is the Gamma function, and $0 < \beta \leq 1$. This formula for the recurrence time distribution can be mathematically derived from the assumption that the probability of a word depends only on the time since its last occurrence, it jumps up whenever the word is used, and decays subsequently by a power-law memory decay function (as proposed in Anderson and Milson 1989, for human memory in general). The stretched exponential (2) is more skewed than the simple exponential distribution (1), which corresponds to the limiting case $\beta = 1$, but less skewed than a power law, which is approached for $\beta \to 0$. This means that the single parameter β effectively captures the burstiness of each individual word. It has low values (close to 0) if the word is extremely bursty, and approaches 1.0 as the word approaches the baseline prediction of the Bag of Words model. Values of $\beta > 1.0$, representing the case in which the word is more evenly distributed than in the baseline prediction, are possible but rarely observed.

Figure 4 shows the interaction of word frequency and word type as predictors of burstiness. At each frequency, the median β values split apart by semantic type. Although there is a visible trend for high-frequency words (low-$\langle\tau\rangle$ words) to be more evenly distributed, the trend associated with semantic class differentiates the β values better.

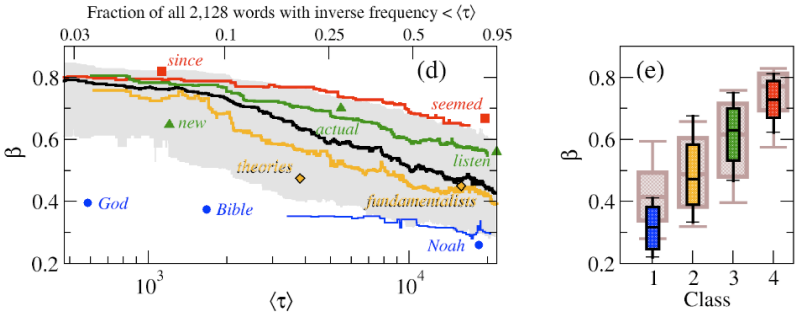

Fig. 4 Relationship of semantic class to β values for words. Left panel: Relationship of $\langle\tau\rangle$ to β. The relationship over all the words is displayed through a running median as a black line. The running median for each of the semantic classes is displayed with a colored line. Example words for each class shown in the same color code. Right panel: Colored boxplots display the distribution of β values for each semantic class. Shadowed boxes were constructed by first ranking words by their frequency, and then binning them in groups that match the semantic class bins in size. The least frequent words are matched to the low β Class 1 nouns, and more frequent words are matched to higher β, higher class nouns. Reproduced from Altmann et al. (2009)

In summary, word type is a stronger statistical predictor of burstiness than a previously reported factor, word frequency. Words of high type (Class 4) tend to be less bursty than words of low type (Class 1), with the other classes falling accordingly in between. Overall, the analysis provides strong quantitative support for a syntagmatic reinterpretation of von Fintel's conjecture.

4 The Behavior of Deverbal Nouns

In Fig. 4, Class 2 is the most numerous class, and it displays the most diversity in β values. This diversity might have arisen because the coarse-grained semantic coding used in Altmann et al. (2009) ignored important semantic distinctions. But it could also arise from factors at the discourse level. Despite the size and length of talk.origins, it still represents a very particular topic, in comparison to the entire space of human discussion. In talk.origins, the word *flood* is very bursty ($\beta = .28$) in comparison to the word *moment* ($\beta = .66$), because of the repeated importance of Noah's flood as an example where biblical and scientific accounts come into conflict. In a physics discussion group, *moment* might be much more bursty, due to its connection with the theory of inertia. Here, I further explore the causes of diversity within the set of common nouns. I consider the behavior of abstract nouns such as *belief*, *argument*, *failure* that are derived from class 3 verbs (*believe*, *argue*, *fail*), in comparison to that of non-derived nouns.

The verb stems of the target nouns have relatively high semantic types because they cannot be defined in terms of sets of properties, but only in terms of functions over predicates or relations. A verb such as *fail* is in this class because of its intensional meaning. It is not possible to determine from direct observation whether someone has failed; for example, falling down is an instance of failing only if the result is contrary to the state of affairs that a person desired. For a clown in a circus, falling down might mean succeeding. A verb such as *evolve* is in this class because it contains an implicit comparison along some scale of time and sets of properties. The nouns derived from these verbs are of interest because they inherit many aspects of the verbal argument structure and semantics. The nouns refer to kinds of activities and events, and full sentences can often be tightly paraphrased as complex noun phrases (Chomsky 1970).

(4) They discussed the theory. Their discussion of the theory.

(5) God created the world. God's creation of the world.

(6) Dawkins asserted that … The assertion by Dawkins that …

The inheritance of the argument structure from the verb stem might suggest that the nouns also inherit intensionality or an implicit scalar comparison from the verb, leading to high logicality, and hence high permutability. If this were true, we would predict systematic differences in burstiness between these abstract nouns, and non-derived nouns of the same frequency. We would predict that these abstract nouns

would share the burstiness of their verbal stems, because they share core components of the stems' semantic structures.

However, for the sentential constructions in (4), the subject is obligatory in English. In contrast, for the nominalized form of these constructions, the expression of the agent (through use of the genitive) is optional, and for some words, it is indeed quite unusual. When using the nominalized form, the speaker might effectively step away from the intensional component of the meaning of the verb. These facts cast doubt on the above predictions because they soften the putative distinction between the derived nouns and other nouns. Further, the similarity between sentential and a nominalized formulations of the same general idea begins to break down if we look at the discourse context. By reifying events and actions, nominalization sets the stage for pronominalization with *it*. Compare (7) and (8):

(7) Their discussion$_i$ of the theory was incomprehensible. It$_i$ included ...

(8) They discussed the theory$_i$ incomprehensibly. It$_i$ included ...

In (7), *It* refers to the discussion of the theory. In (8), *It* more readily refers to the theory. This observation can be connected to Carlson's treatment of topic (Carlson 1983), which very innovatively treats topics as questions under discussion. Because nominalized constructions can be used to instantiate *wh* question pronouns, they can be used to answer questions in straightforward way, as in (9). In the same context, the sentential construction is less felicitous; compare (9) and (10).

(9) What was incomprehensible? – Their discussion of the theory.

(10) What was incomprehensible? – #They discussed the theory.

Example (10) is less felicitous than (9) because it involves a bridging inference; the two sentences are related only via the implict assumption that events can cause mental reactions. Such indirect connections between discourse moves, already prefigured in Carlson (1983), have been subsequently found to place demands on working memory and to vary across individuals (Singer et al. 1992).

4.1 Materials

In the present study, words are defined very superfically as strings of alphabetic characters separated from other strings by white space. Space, tab, and newline, underscore, and punctuation marks *: ; . ! , ?* are treated as white space. In computational linguistics, words with the same stem but different inflectional endings are often collapsed together using a lemmatization algorithm. For example, the singular and plural forms of a noun, or the present and past forms of verbs, are often grouped together with a view to increasing the sample size for the stem. No lemmatization was used in the present study, because the sample sizes are extremely ample, morphologically related words often differ in their contexts of use, and one goal of the study was to understand how they differ.

Specifically, the study looks at the 43 pairs of Class 3 verbs and deverbal nouns for which both members occurred 10,000 times or more in the talk.origins dataset. The nouns were formed with a variety of deverbal suffixes: *-ion*, *-ment*, *-tion*, *-ure*, *-al*, *-er*, *-th* (cf. *evolve/evolution*, *argue/argument*, *fail/failure*, *survive/survival*, *teach/teacher*, *grow/growth*). In order to avoid statistical dependence amongst different word pairs, only one pair of words was used for any given stem, even if several pairs were found in the data. If both the singular and plural form of the noun met the frequency threshhold, only the singular form was used. If a verb occurred both in its bare form (corresponding to the 1st and 2nd person present and the infinitive) and in an *-ed* form (corresponding to the past), the bare form was used.

A control set of non-derived nouns was also selected. Because frequency is a known factor in burstiness, these were frequency-matched to the derived nouns. For each derived noun, the noun closest to it in frequency was taken. To match the selection rule for deverbal nouns, plural forms were used only when no nearby singular form was available. Duplication of control items was avoided by taking the second-closest frequency match in the few cases in which the same non-derived word fell closest to two of the derived nouns. Control words included both concrete nouns, such as *book*, *fossil*, *apes*, *child*, *house* and abstract nouns such as *science*, *number*, *context*, *structures*. The argument structures of the nouns played no role in their selection, and many of them can in principle take arguments: cf. *a book about Darwin*; *a dinosaur fossil*; *the House of God*; *the number of species*. . . .

4.2 Results

The control nouns are more bursty than the verb stems. This is evident from their significantly lower β values in the two left boxes of the boxplots in Fig. 5. This result is expected, reproducing within a subset of the data one of the general findings displayed in the previous section. The distribution of β for derived nouns, as shown at the right, essentially matches that of control nouns. In the aggregate, the derived nouns do not inherit the burstiness of Class 3 verbs. The morphological derivation reduces the β value (increasing the burstiness), to such a complete extent that the derived nouns behave like any other noun.

The relative burstiness of the derived nouns in relation to the two comparison sets is displayed in a different way, by the individual word pairs, in Figs. 6 and 7. Figure 6 displays the comparison between control nouns and derived nouns. The points are evenly distributed around the diagonal comparison line, indicating that morphological complexity is not associated with any systematic pattern.

A similar comparison is made in Fig. 7 for the verb stems (e.g. *argue*) in relation to the derived nouns (e.g. *argument*). For all but two word pairs, the verb has a higher β value than the noun.

These results indicate that the morphological derivation very systematically reduced the β value associated with the verbal concept when converting it into a noun. However, the detail in Fig. 7 also reveals some correlation between the verb bursti-

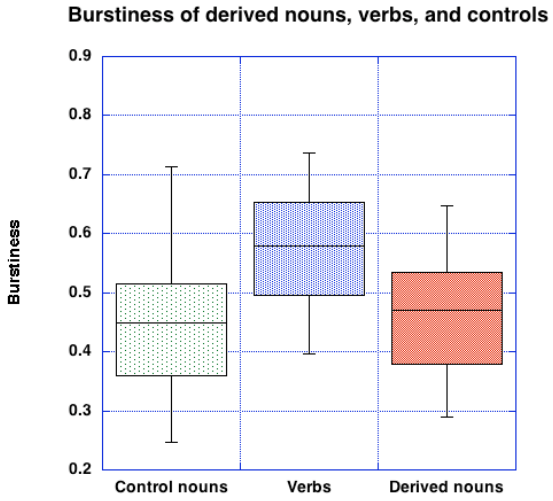

Fig. 5 Distribution of β values for control nouns (such as *science*), verbs (such as *argue* , and derived nouns (such as *argument*). Boxplots defined as the median, quartiles, and octiles of the

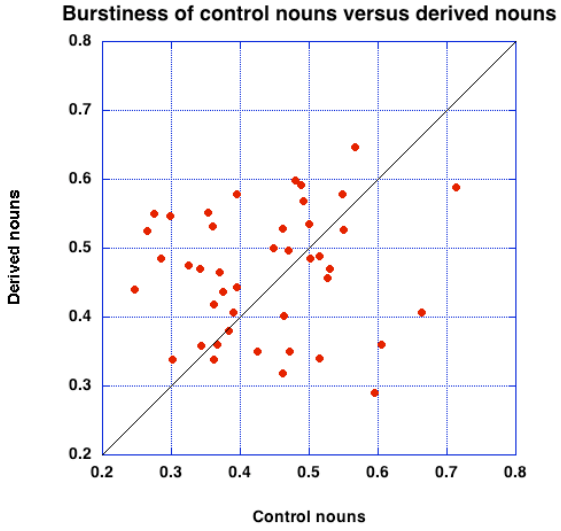

Fig. 6 Burstiness of control nouns versus derived nouns. Each plotting point represents the relationship of the β value for a derived noun (e.g. *argument*) to its frequency-matched control noun (e.g. *science*). The diagonal line $y = x$ shows where the points would fall if these values were equal

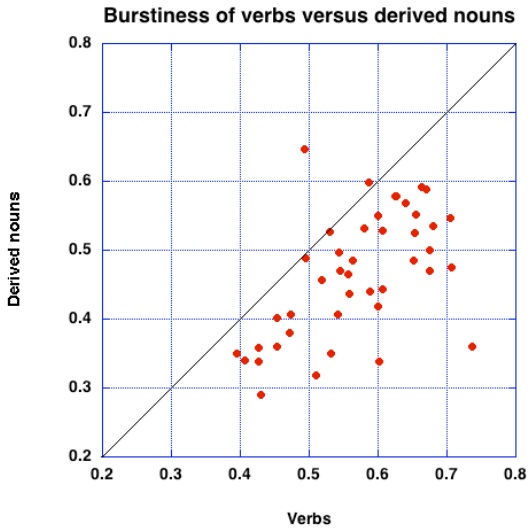

Fig. 7 Burstiness of verbs versus derived nouns. Each plotting point represents the relationship of the β value for a derived noun to its stem verb. The diagonal line $y = x$ shows where the points would fall if these values were equal

ness and that of the derived noun, that was not evident from the boxplots. Extremely bursty verbs, such as *evolve*, *predict*, *survive*, *measured* and *teach* have even more bursty nouns. Less bursty verbs, such as *argue*, *refer* and *ignore* also tend to have less bursty nouns. Overall, the most bursty noun-verb pairs do not seem to differ from the others in their logical structure. So, we speculate that this correlation arises through more general associative mechanisms. Church (2000) already showed that the occurrence of a topical word in a text increases not just probability of the same word, but also the probabilities of its semantic associates. Morphological relatives are also semantic associates, and therefore should share patterns of distribution in the text. For example, *evolve* and *evolution* are highly topical in this dataset through their connection with Darwinian theory, and should therefore tend to occur in the same bursts. *Teach* and *teacher* may be more associated with the creationist point of view, and so occur in other bursts. *Argument* is widely applicable to different topics and points of view, and *argue* shares this property.

5 Discussion and Conclusion

The tendency of topical words to reoccur in bursts is a mainstay of document indexing and retrieval. However, the relationship of burstiness to the structure of the linguistic system has been little explored. Here, I have reviewed a formal apparatus

for elucidating burstiness patterns that was developed in Altmann et al. (2009). Using the single free parameter β of the stretched exponential (Weibull) distribution to parameterize word recurrence distributions, it was possible to establish a connection between burstiness and semantic type. Overall, words of high semantic type prove to have recurrence distributions that are much closer than lower type words to what would be expected under a Bag of Words model, in which the words in a text are simply assembled in random order. This supports the three-way association of semantic type, logicality, and permutability that was advanced in von Fintel's (1995) conjecture.

The same apparatus was also applied in a novel exploration of derivational morphology. A set of nouns derived from intensional verbs by the addition of a suffix was compared both to the corresponding set of verbs, and to a control set of non-derived nouns. The primary empirical findings were:

- The distribution of burstiness values for the derived nouns is systematically lower than for the verb stems.
- The distribution of burstiness values for the derived nouns matches that of the control nouns.
- In a paired comparison, there is a correlation between the burstiness of the verb stem and the burstiness of the corresponding derived noun.

These three observations can be integrated by assuming that the deverbal suffixes exemplified in this study (*-tion*, *-ment*, *-al*, *-ance*, etc.) lower the semantic type of the stem they attach to. However, this formal operation does not affect the associative structure of the lexicon, with the result that noun-verb pairs with shared thematic or social connotations can exhibit correlated burstiness values.

Why aren't the deverbal nouns more permutable? One possibility is that they have lost the semantic type features of their stem verbs. For example, we can compare the usage of *evolve* and *evolution*. *Evolve* requires a specification of a lineage, and the lineage varies from one type of organism to another. Modern birds *evolved* from dinosaurs, whereas modern ferns *evolved* from earlier plants. Supporting von Fintel's intuition, *evolve* is equally a propos in a discussion about birds and a discussion about dinosaurs. The same could be said about the word *evolution*, except that *evolution* has acquired a further sense in which it steps back from questions about how anything in particular evolves, instead referring to the specific theory claiming that all living things evolve. The existence or veracity of this theory can then serve as a topic of discussion in itself. Similar observations can be made about deverbal nouns such as *measurement* and *direction*. These lack the intensional baggage of the stem in many common uses, such as *the wrong direction* or *his wrist measurement*. From a semantic point of view, the deverbal affixes in this study can thus be considered as providing type-lowering. In Altmann et al. (2009), the deadjectival affix *-ly* (as in *frequent*, *frequently*) was similarly found to decrease the burstiness associated with its stem, and by inference to raise the semantic type.

This observation can be connected to the theoretical discussion surrounding the concept of a *head* in syntax and morphology. The head-dependency relations in a complex form (whether a morphologically complex word, or a syntactic phrase)

control how properties of the parts contribute to the properties of the whole. The verb stem might be construed as the head in of a nominalization because its lexical semantics (including its overt or implied argument structure) characterizes the sort of event that the nominalization refers to. However, the nominalizing suffix meets the technical and morphosyntactic definitions of *head* presented in Hoeksema (1992); the suffix determines the category of the nominalization and is the locus of inflection. I have here identified a novel correlate of these formal properties, namely statistical signatures at long, discourse-level, timescales. The tendency to drop arguments in nominalizations (compared to the maximal argument structure that could be supported) may be viewed as a further reflex of the dominance of the suffix over the stem.

There is a tension between this suggestion and the idea that semantic type is strongly correlated with logicality and permutability. The term *logicality* was defined as the extent to which the meaning of a word is immune to specific facts about the world. The connection of high type to logicality and permutability follows from the assumption that any relation can be instantiated with a wide variety of entities, and that any function over relations can be instantiated with a wide variety of relations. According to this thinking, relations would inherit all the different contexts of use from the different entities that can instantiate them, and functions over relations would doubly benefit, by inheriting contexts of use from all the different relations. However, intuitively, the claim that *measurement* has lower logicality than *measure* (and hence lower permutability) does not seem right, since anything that can be measured has the corresponding measurement.

This tension reveals an important hidden assumption in the original proposal. The argument would go through if people's conceptualization of reality used a uniform and fixed level of granularity. But it doesn't. A noteworthy cognitive capability of humans is the ability to ramify concepts, that is to elaborate them by taking up more and more questions about them. Though these informational elaborations may involve powerful abstract relations, with a variety of alternative instantiations, they do not necessarily encompass any greater fraction of reality than before. For example, whereas a small child considers *dance* to be an undifferentiated activity associated with music, almost any Finnish adult would differentiate tangos, waltzes, and foxtrots, and other dances. They can build on this variety to acquire the abstract verbal concept *to syncopate*. Although this concept is abstract, and can be instantiated in a variety of different beats, its empirical applicability still does not extend outside of the world of music and dance. Within this world, it has high permutability; but within the world at large, it does not.

In Carlson (1983), discourse is a game in which questions provide the basis for the construction and elaboration of shared knowledge. In this framework, the decision to use a nominalized construction, in preference to a nearly synonymous sentential construction, can be viewed as a game move. Instead of thinking about lowering the semantic type for the stem, we can instead think about raising the point of view of the discussion. Dennett and Haugeland (1987) sketch a theory of *intentionality* in discourse, in which speakers communicative choices reflect their *intentions* of pointing at whatever they are speaking *about*. The choice of a deverbal

noun is *intentional* because it reflects the speaker's intention to point at an event or relation. By doing so, the speaker opens the way to further ramification of the information associated with that event or relation. Recalling that the ramification of topics causes the burstiness of topical words, this account provides a mechanism for deverbal nouns to be as bursty as other nouns. Unlike the type-lowering account, the mechanism would apply whether the deverbal noun is bleached of its intensional baggage or not. One can indeed conjecture that type-lowering might be a long term consequence of using a deverbal noun to raise the point of view repeatedly, in a variety of different contexts. This diachronic trajectory would be consistent with other cases in which pragmatic choices eventually become encapsulated in the semantics of a language.

References

Altmann, Eduardo G., Janet B. Pierrehumbert, and Adilson E. Motter. 2009. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS ONE* 4:e7678. doi:10.1371/journal.pone.0007678.

Altmann, Eduardo G., Janet B. Pierrehumbert, and Adilson E. Motter. 2011. Niche as a determinant of word fate in online groups. *PLoS ONE* 6:e19009. doi:10.1371/journal.pone.0019009.

Anderson, John R., and Robert Milson. 1989. Human memory: An adaptive perspective. *Psychological Review* 96:703–719.

Baayen, R. H., Lee H. Wurm, and Joanna Aycock. 2007. Lexical dynamics for low-frequency complex words: a regression study across tasks and modalities. *The Mental Lexicon* 2:419–463. doi:10.1075/ml.2.3.06baa.

van Benthem, Johan. 1989. Logical constants across varying types. *Notre Dame Journal of Formal Logic* 30:315–342.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Blevins, Juliette, and Andrew Wedel. 2009. Inhibited sound change: An evolutionary approach to lexical competition. *Diachronica* 26:143–183. doi:10.1075/dia.26.2.01ble.

Bookstein, Abraham, and Don R. Swanson. 1974. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science* 25:312–318. doi:10.1002/asi.4630250505.

Bybee, Joan. 2001. *Phonology and language use*. Number 94 in Cambridge Studies in Linguistics. Cambridge: Cambridge University Press.

Carlson, Lauri. 1983. *Dialogue games: An approach to discourse analysis*. Number 17 in Synthese language library. Dordrecht: Reidel.

Chomsky, Noam. 1970. Remarks on nominalizations. In *Readings in English transformational grammar*, ed. Roderick A. Jacobs and Peter S. Rosenbaum, 184–221. Waltham MA: Ginn &Co.

Church, Kenneth W. 2000. Empirical estimates of adaptation: The chance of two Noriegas is closer to $p/2$ than p^2. In *Proceedings of the 17th conference on Computational linguistics (COLING 2000)*, 180–186. Association for Computational Linguistics.

Church, Kenneth W., and William A. Gale. 1995. Poisson mixtures. *Natural Language Engineering* 1:163–190. doi:10.1017/S1351324900000139.

Dennett, Daniel C., and John Haugeland. 1987. Intentionality. In *The Oxford companion to the mind*, ed. Richard L. Gregory, 383–386. Oxford University Press.

von Fintel, Kai. 1995. The formal semantics of grammaticalization. In *Proceedings of NELS 25. Volume 2: Papers from the Workshops on Language Acquisition & Language Change GLSA*, 175–189.

Hay, Jennifer. 2003. *Causes and consequences of word structure*. Routledge.

Heller, Jordana, and Janet B. Pierrehumbert. 2011. Word burstiness improves models of word reduction in spontaneous speech. In *Architectures and Mechanisms for Language Processing (AMLaP 2011)*. Paris. http://amlap2011.files.wordpress.com/2011/08/129_pdf.pdf.

Heller, Jordana, Janet B. Pierrehumbert, and David N. Rapp. 2010. Predicting words beyond the syntactic horizon: Word recurrence distributions modulate on-line long-distance lexical predictability. In *Architectures and Mechanisms for Language Processing (AMLaP 2010)*. York, UK: University of York.

Hoeksema, Jack. 1992. The head parameter in morphology and syntax. In *Language and cognition 2: Yearbook 1992 of the research group for Linguistic Theory and Knowledge Representation of the University of Groningen*, ed. Dicky Gilbers and Sietze Looyenga, 119–132. Groningen: Universiteitsdrukkerij Groningen.

Katz, Slava M. 1996. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering* 2:15–59.

Kintsch, Walter. 1974. *The representation of meaning in memory*. The experimental psychology series. Hillsdale, NJ: Erlbaum.

Lijffijt, Jefrey, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila. 2011. Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping. In *Machine Learning and Knowledge Discovery in Databases. European Conference, ECML PKDD 2011. Proceedings, Part II*, ed. Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, number 6912 in Lecture Notes in Artificial Intelligence, 341–357. Berlin, Heidelberg: Springer.

Montague, Richard. 1973. The proper treatment of quantification in ordinary English. In *Approaches to natural language*, ed. Jaakko Hintikka, Julius Moravscik, and Patrick Suppes, 221–242. Dordrecht: Reidel.

Montemurro, Marcelo A., and Damián H. Zanette. 2002. Entropic analysis of the role of words in literary texts. *Advances in Complex Systems* 5:7–17.

Nigam, Kamal, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning* 39:103–134. doi:10.1023/A:1007692713085.

Partee, Barbara H. 1992. Syntactic categories and semantic type. In *Computational linguistics and formal semantics*, ed. Michael Rosner and Roderick Johnson, Studies in Natural Language Processing, 97–126. Cambridge, UK: Cambridge University Press.

Sarkar, Avik, Paul Garthwaite, and Anne de Roeck. 2005. A Bayesian mixture model for term re-occurrence and burstiness. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL)*, 48–55.

Sharkey, Noel E., and D. C. Mitchell. 1985. Word recognition in a functional context: The use of scripts in reading. *Journal of Memory and Language* 24:253–270. doi:10.1016/0749-596X(85)90027-0.

Singer, Murray, Peter Andruslak, Paul Reisdorf, and Nancy L. Black. 1992. Individual differences in bridging inference processes. *Memory & Cognition* 20:539–548. doi:10.3758/BF03199586.

Tanenhaus, Michael K, and Sarah Brown-Schmidt. 2008. Language processing in the natural world. *Philosophical Transactions of the Royal Society B* 363:1105–1122.