Poznań Studies in Contemporary Linguistics 46(2), 2010, pp.193–219

doi:10.2478/v10010-010-0010-x

PRODUCTIVITY VS. LEXICALIZATION: FREQUENCY-BASED HYPOTHESES ON WORD-FORMATION

JESÚS FERNÁNDEZ-DOMÍNGUEZ
University of Jaén
jesusferdom@gmail.com

ABSTRACT

This article looks at *morphological productivity* and *lexicalization*. Productivity, first, bears a significant relationship with frequency because both seem to be subtly interlinked through low-frequency items. Much the same happens between lexicalization and frequency, although their association must be seen from a different angle because lexicalized words tend to have greater frequencies than non-lexicalized words. The novelty of this paper is that it provides a link between the above two notions and corpus-based frequency figures, and then operates a formula (π) on two sets of units, some lexicalized, some synchronically analysable. The two subcorpora confirm a correct function of π to tell between words which tend to be used by means of word-formation vs. words which already exist in the individual's lexicon.

KEYWORDS: Corpus linguistics; lexicalization; morphology; productivity.

1. Introduction

Much has been discussed as far as the two concepts presented in this article are concerned. Morphological productivity, on the one hand, is gradually recovering from the period of confusion and disorientation where it has been trapped for a substantial amount of time (cf. Bauer 1983: 62; see Bauer 2005 for an overview of the state of the art), and not few works concerned with this phenomenon are currently at hand, some attempting its quantification (Copestake 2001; Baayen 2005), some paying heed predominantly to theorizing on it (Bauer 2001, 2005).

The other notion examined here is lexicalization. As opposed to the results of synchronic word-formation, lexicalization is concerned with units that are stored in the lexicon and which, roughly speaking, have to be memorized for use (Plag 1999: 9–11; Jackendoff 2002: 156). It follows that productivity and lexicalization represent the two possible sources for the fulfilment of a naming need and, although not directly compa-

rable on the same scale, both are at hand for users to pick them up when required (this is explained later in 3).

The fact is that, even if occasional pieces of research have associated both notions (Aronoff 1983; Aronoff and Anshen 2001; Bauer 2004), there seems to be an implicit relationship between productivity and lexicalization to be discovered yet. This paper aims to explore such areas, paying special attention to the following matters:

(i) the tendency for low frequency in productivity vs. high frequency in lexicalization;
(ii) a dynamic model of word-formation which can account for such differences in frequency by proposing two well-defined routes (one for word-formation, one for the lexicon);
(iii) a corpus-based formula (π) to provide a numerical justification for profitability in synchronically-analysable items and for lexicalization in stored items.

To this end, a 5,878-noun corpus was derived from the *British National Corpus Sampler* (hereafter, *BNC Sampler*) on which a number of operations were carried out. These confirm that the status of words as lexicalized or not is robustly linked to their frequency which, in turn, can be matched to the frequency of use of one of the two possible routes by speakers (these are discussed in Section 2.1).

The distribution of contents in this paper is as follows: Section 2 immerses into productivity and lexicalization in theoretical terms, while Section 3 pays attention to the compilation of the study corpus. The data-based experiments are carried out in Section 4, and Section 5 is the conclusions of the research.

2. Productivity and lexicalization

The two notions in the title of this section have occupied linguists for the past twenty years due to their importance for word-formation, although the attention which they have attracted is paralleled by the difficulties of their definitions. This section provides an up-to-date recapitulation of both notions and sets up the foundations for the experiments in Section 4.

2.1. Productive word-formation

Morphological productivity is an area of research that has become the focus of attention in recent years, being approached from different angles both in the theory and in the practice. It proves hence virtually impossible to provide an authoritative definition of productivity without ignoring one part of the literature or another, as there are aspects of it that are perceived as indispensable by some authors but are denied by others, and vice versa, e.g. the role played by frequency, a possible differentiation between existing and

potential words, or the relationship between transparency and opacity (see Aronoff 1983; Bauer 2001: 33–99; Štekauer 2005: 226–230). In essence, a word-formation process is said to be productive if it has the potential for speakers to operate it in an unconscious and repetitive way for the rule-governed production of an indefinite number of words (see Schultink 1961: 113; Aronoff 1983; Plag 1999: 11; Bauer 2001: 211). What follows is an outline of a dynamic system of productivity and its relationship to concepts like the *lexicon*, *lexicalization* or *frequency*.

Customarily, descriptions of morphological productivity portray it as a knotty and problematic phenomenon, undoubtedly due to the complexities which it still hides. But a part of these difficulties surely comes from past misunderstanding because, although traditionally overlooked, productivity is in fact a two-sided phenomenon, decomposable into *availability* and *profitability* (see Corbin 1987: 177; Carstairs-McCarthy 1992: 37; Bauer 2001: 48–49). Availability has to do with whether a process can be used at a certain moment or not, and is thus a yes/no question because a process is either available or it is not: it is a qualitative view of productivity. Profitability, on the other hand, is a quantitative notion because it deals with how many lexemes an available process coins, hence one process can be *more* profitable than other.

Earlier scholars depicted productivity as a whole, probably unaware of its two hyponyms, but the fact is that dismissing availability and profitability in contemporary linguistics means neglecting part of the productive potential of a morphology. The burden of past confusion, however, falls mainly on linguists studying this phenomenon today, because an only term, *productivity*, encompasses so many notions that it is difficult to actually discriminate different uses of it. Leaving terminological inconsistencies aside, the division availability vs. profitability has become a milestone in word-formation, and it should be given appropriate weight in contemporary morphological models.

The starting point of the system presented here is naming needs, the force that sets word-formation in motion and without which the act of designating is unnecessary. Parallel to other views (Bauer 2001; Štekauer 2005), this model is founded on the belief that no word can be productively[1] coined if it is not required by the community. In this approach, a perceived need may result in either the picking up of a unit from the individual's mental lexicon, or in the creation of a new word; the former is relevant to lexicalization, the latter to productivity. This is the idea presented by Figure 1.[2]

---

[1] For other types of coinages, the label *creativity* has been suggested. *Creativity* represents formations that do not necessarily indicate the productive use of a word-formation process, as in words from literary texts, individual inventions, playful formations or new technical terms (see Bauer 1983: 63–64, 2005: 330; Lipka 2002: 92; Nishiwaka 2003: 37–38).

[2] This rationale complies with the tenets of the dual-route models, as founded by Pinker and Prince (1988; see Pinker 1991), who propose a theory of inflectional morphology constituted by two mental routes, one being rule-governed and relevant to on-line formations, the other being a memory system of stored items. In that theory, the regular pattern that occurs most is set as the default rule, e.g. /ed/ in past-tense formations, while irregular constructions have to be memorized (see also Rumelhart and McClelland 1987; Marcus et al.

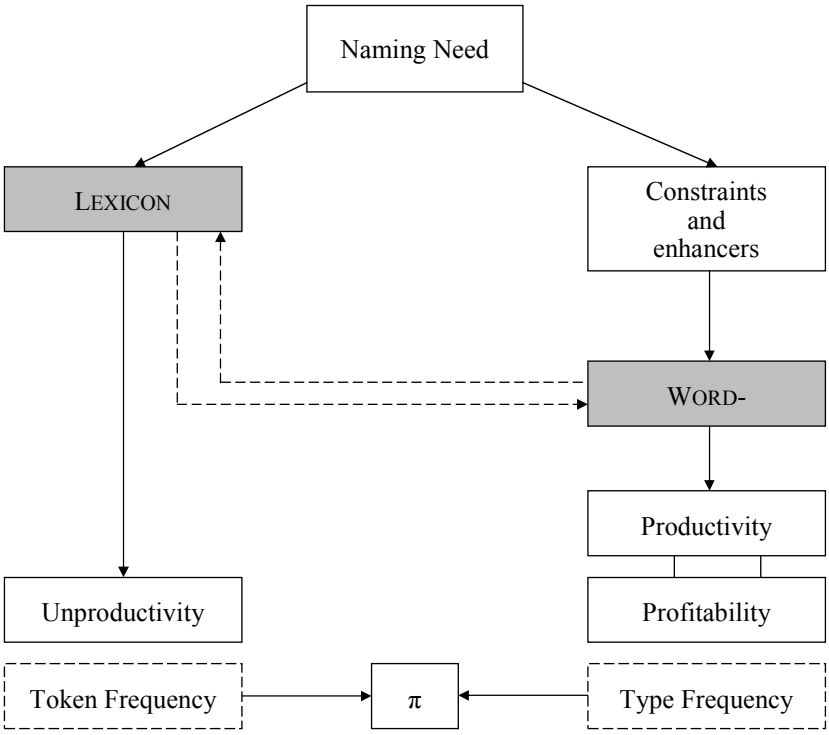

Figure 1. A dynamic model of productivity and lexicalization.

The above mechanism, then, is initiated with the scanning of the mental lexicon in search for a fitting word as the first reaction to a need. Here, as previously explained, two possibilities exist. One is that the lexicon contains a suitable lexeme, in which case the entry is automatically selected and generated as output.

The reason why this stage requires no further operation is that units listed in the lexicon often retain grammatical information with them (e.g. morphological, semantic, syntactic, etc.), so only their uttering remains (Allan 2006: 148). Note that this *lexicon route* does not have a bearing on morphological productivity, because the word-formation module is not involved in its production.

This path, however, does affect the word's token frequency (hereafter *N*). As it is shown in 2.2, the relevance of a lexeme for the speaking community at a given time is revealed by common occurrence, which in turn leads to a higher frequency. In the case of a word like *dog*, the lexicon route is repeatedly activated to express a notion through a specific lexeme, and this has eventually turned the word into an institutionalized one.

1992; De Jong et al. 2000; Hay 2003; Hay and Baayen 2005). The model here proposed applies these thoughts to the area of derivational morphology.

The lexeme *dog* is regularly employed in language and, as a result, it has a high *N* in a corpus like the *BNC Sampler*: 468. The figure of *N*, then, reflects the incidence of usage of the lexicon route so, the higher the token frequency of a lexeme the more often the lexicon route has been used for it.

The second possibility in Figure 1 is that the speaker's naming need is not mirrored by any unit in the lexicon, in which case the *word-formation route* is taken. In this case, a new meaning has to be conveyed, so a new word needs to be created because "[i]t is impossible to assign a morphosyntactic specification to a meaning without also assigning it a form" (Allan 2006: 150). Note that the lexicon route has priority over the word-formation route so that, when the former can lead to successful output, the latter cannot be taken (this has been studied under the constraint of *blocking*, see Aronoff 1976: 43–53; Plag 1999: 50–54; Bauer 2001: 136–139; Dressler 2007: 168).

The word-formation route starts with the potential derivative having to overcome specific enhancers and constrains, which may either favour or limit its production. There are many factors that affect processes in this manner, like blocking, hypostatization[3] or constraints concerned with morphology, syntax, semantics and pragmatics at different levels (see Lipka 1977: 161; Bauer 1983: 84–99, 2001: 126–143; Plag 1999: 37–61). Constraints and enhancers mean a crucial stage because they may restrain lexical bases, the combinability of affixes, the processing of units or the input/output of the process in question so, when a derivative succeeds in passing through them, it achieves productive output (see Kuperman et al. 2008; Plag and Baayen 2009).

Subsequently, word-formation needs to nourish from the lexicon to obtain the lexical base(s) for derivation, therefore both modules have to be bidirectionally connected (hence the dashed arrows linking them in Figure 1). This happens because, while the lexicon results from the accumulation of records, word-formation is a dynamic and ever-changing component containing only synchronically accessible processes. The word-formation module, thus, has to obtain lexical bases from an external place for one of the available rules to generate a new word. For example, for the production of *happily* the lexical base *happy* would have to be captured from the lexicon, and then derivation with *-ly* applied by word-formation. This view implies the previous existence of a lexeme for the creation of a new one because, if no base exists in the lexicon, no derivative can be produced[4] (see Schultink 1961: 113; Carstairs-McCarthy 1992: 25–27; Nishiwaka 2003: 38).

Given the dynamic features of the word-formation component, an essential aspect is that solely those processes with the feature [+AVAILABLE] may form part of it. This at-

[3] Coined by Lipka (1977: 161) originally as German *Hypostasierung*, this restriction refers to the fact that the production of a word implies the (mental) existence of the object which it denotes and, if strictly applied, it implies that only existing things can be labelled.

[4] The term *word*-formation *process* is not considered here a synonym for *affixation*, therefore compounding (*sail* and *boat* > *sailboat*) and conversion ($plane_N$ > $plane_V$) are considered to require a base for derivation too. The same applies to minor word-formation processes like back-formation (*bulldozer* > *bulldoze*) and blending (*cybernetic* and *organism* > *cyborg*).

tribute guarantees that synchronically accessible rules are the only ones to be employed for derivation and entails that, if a process loses this attribute, it leaves this module for the lexicon, where it can serve other purposes (see Chomsky 1964, 1965; Lipka 1977; van Marle 1985). The relevance of the feature [±AVAILABLE] lies not only in placing special emphasis on the quantitative vs. qualitative side of productivity, but also in acting as a filter for productive creations, such that the output from unavailable processes falls under the domain of creativity and not of productivity. The feature [±AVAILABLE], in short, prevents synchronically unavailable word-formation processes from coining lexemes, and builds a link with the distinction *Wortbildung* vs. *Wortgebildetheit*, started by Dokulil (1962; see Stepanova 1973; Aronoff 1976; Dressler 2003).

Once the speaker has unconsciously chosen the lexical base(s) and the derivative has been generated, productive output is achieved and, depending on how often a word-formation process is exploited, it will have a higher or lower profitability rate, i.e. it will have a specific degree of productivity. In this model, application of the word-formation route happens once per derivative because, once a coinage has been created, it may be used repeatedly, but it will never pass through the word-formation component again (i.e. it will never be created again), hence the label *once-only rules* by Aronoff (1976: 30) and Spencer's remark that "once a word has been formed [...] it can't be unformed" (1991: 84). Further uses of a coined lexeme will involve its repetition as an existing word but not its creation from scratch (see Section 2.2).

As happens with the lexicon, the word-formation route has an effect on frequency. On this occasion, because word-formation is an only-once procedure, it is type frequency (hereafter, *V*) that becomes influenced. We know that types are the number of different words that have been coined by means of a process, which means that, for a given rule, one type corresponds to one derivative, two types symbolize two derivatives, and so on (see Plag 1999: 27). If so, and accepting the assumption that corpora are reliable reflections of language (Bauer 2001: 47; Plag 2003: 52), *V* should be a good indicator of the number of words coined by a process, so that the higher the figure of types, the more units a process has formed.

As has been shown in this section, the figures of type and token frequency can be associated with each of the two possible routes in the proposed system (Figure 1), which carries certain implications for word-formation processes. In section 4, we will return to this issue and argue in favour of the profitability-measuring formula π as a pattern to compute the frequency of use of the two routes.

In the next section, the question of lexicon-based frequency is further discussed together with lexicalization, the usefulness of words and token frequency.

2.2. Lexicalization and the lexicon

Lexicalization is one of the possible stages in a word's life, occurring at an advanced point of it (if at all). Figure 2 shows three of those possible phases, the first of them being its coinage, that is, the production of a word which did not exist before.

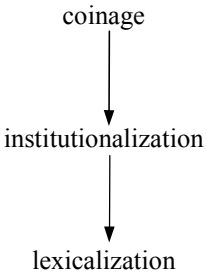

Figure 2. Stages in a word's life.

A unit produced at the first stage can also been called *ad-hoc formation*, *neologism* or *nonce-formation* (Bauer 1983: 45; Hohenhaus 2005: 363), and takes place if a speaker scans the mental lexicon without finding a suitable lexeme for the required meaning. This is an unavoidable stage in the life of existing words, although previous steps have been hinted where words are invisible to the community.[5]

An important feature of coinages is that they may be produced and be used for some time, but they may as well disappear without a trace and their lifespan will come to an end. The existence of those words, given their briefness, will have been manifest only to the speakers who coined them and perhaps for the speaking community around, but they will go unnoticed for most language users (Lipka 2002: 112; Hohenhaus 2005: 360–361). A coinage may, alternatively, be produced and widely employed for a reasonable amount of time. In that case, the word becomes *institutionalized*, i.e. "the nonce formation starts to be accepted by other speakers as a known lexical item" (Bauer 1983: 48). The lexeme at this stage is recognized by language users as one more item of their regular vocabulary, so that its formal and semantic features are not seen as salient or deviant. However, the shift from coinage to institutionalized word can be hardly pinned down accurately, as institutionalized words can be unmistakably recognized only once they are at this stage.

One basic distinction between coinages and institutionalized words lies at their semantics: a potential word may have various possible senses, only one of which is selected upon coinage depending on the precise needs. Take, for instance, the potential compound *rain-snake* and several of its potential meanings (see Bauer and Huddleston 2002: 1647):

---

[5] Words are *possible*/*potential* if they can be readily produced but have not yet; put differently, existing words are a subset of the possible. The notion of potential word is strongly linked to the generative branch of linguistics and implies an *overgenerating capacity* of word-formation, that is, language is able to create more words than are actually needed. That is why some words are needed and coined, and others have not been needed yet and are waiting to become visible (see Aronoff 1976: 36; Bauer 2001: 40–41; Dal 2003: 12–14; Hohenhaus 2005: 360).

(1a) a snake which comes out in the rain

(1b) a snake made of rain

(1c) a snake which causes rain

One of the possible meanings (a–c) may be picked up if this word is ever coined, and it will accompany the formal structure *rain-snake*. That particular meaning will be kept in subsequent uses of the lexeme, so that the semantic side of the word is progressively associated to its formal make-up; that is, the word has been immediately lexicalized to some extent. After a reasonable time, there may be a point when *rain-snake* becomes institutionalized due to continued use, so that its denotation is perfectly well-known to language users and interpretations other than the original one are rarer. Now, any use of *rain-snake* where its meaning diverges from the standard one will be seen as marginal or peripheral, as in puns or word-plays.

At this point, the institutionalized word may vanish into oblivion if unneeded, or it may go on being used and, in the latter case, two possibilities exist. The first is that the unit remains institutionalized, the situation of *goldsmith* and *stone wall* in Contemporary English; the second possibility is that the lexeme becomes affected by lexicalization, the third phase in Figure 2. Leaving aside senses of the term beyond the aim of this paper,[6] lexicalization has been customarily defined as "the phenomenon that complex lexical items, through frequent usage, may lose their syntagmatic nature and tend to become formal units with specific content" (Lipka 2002: 97). For these purposes, lexicalization is synonymous with *being part of the lexicon*,[7] in that items that are affected by it can no longer be generated by word-formation and have to be necessarily listed.

In the system presented in this paper, the lexicon is a list of units, of whatever origin or level, which cannot be derived by means of synchronic word-formation. It is thus an individualized module that contains words, but also lexicalized phrases or even sentences, although the proportion is higher in the former two than in the latter for reasons of storage size. The lexicon also registers those word-formation processes which are

---

[6] In diachronic studies, lexicalization has sometimes been opposed to grammaticalization, both perceived as processes which make words fluctuate from the lexicon to the grammar and vice versa (see, for example, Brinton and Traugott 2005), although this differentiation is not always observed. A second sense, in the field of Generative Semantics, refers to "a process in which a configuration of semantic elements in an abstract representation is replaced by a lexeme" (Lipka 2002: 111). For more on senses of *lexicalization*, see Sauer (2004: 1625–1627), Hohenhaus (2005: 353–357) and Bakken (2006: 106).

[7] The lexicon (from the Greek *λεξικός* 'dictionary') has often been characterized as "a list of irregularities" (Bloomfield 1933) or as a component made up of simplex items to feed syntax, but several other senses can be distinguished in this term (see Nishiwaka 2003: 32–34; Brinton and Traugott 2005: 9–18; Allan 2006). It has also been understood as made up of only words (Aronoff 1976; Corbin 1987) and as conformed by words plus affixes (Selkirk 1982), although others propose to see it essentially as a store for *listemes*, i.e. any linguistic object that has to be memorized by speakers (Di Sciullo and Williams 1987). There are attributes of the lexicon which vary depending which view we ascribe to, but one aspect is common to them all: only existing items may form part of it. Put simply, "what is not part of it does not exist" (Kastovsky 1986: 586).

unavailable synchronically and cannot be used productively today; it is, in short, a repository for every device that word-formation may need for the creation of coinages (see Section 2.1). A general assumption is that the lexicon should contain as few items as possible, namely idiosyncratic and irregular forms whose features cannot be captured by rules. Items that speakers do not need to remember, hence, should remain outside the lexicon and be regular, instantaneously interpretable by rules upon perception (Verspoor 1998: 152; Štekauer 2000: 10–20, 2001; Jackendoff 2002: 155–158). From the above, it emerges that an appropriate theory of word-formation should not ignore lexicalization given the direct relevance which it has for complex lexemes (see Riehemann 1998: 51–52).

Lexicalization, it has been maintained (Jackendoff 2002; Bauer 2004: 19), is not an either/or concept, because words reveal its application in varying degrees. There are, not in vain, different types of lexicalization depending on the level of description involved therein (phonological, morphological, semantic or syntactic). In these cases, lexicalization is triggered by different factors, respectively: placement of stress on a syllable different than the regular one (here the first syllable instead of the one preceding the suffix) (2a), use of a synchronically unavailable suffix (2b), having an irregular unpredictable meaning (2c) or displaying an internal syntactic structure (2d).

(2a) ˈArabic, ˈchivalric, ˈcholeric
(2b) achievement, assignment, derailment
(2c) mincemeat, pushchair, wheelchair
(2d) pickpocket, scarecrow, telltale
(2e) length, width
(Bauer 1983: 51–61)

Thus, one lexical item may be affected by only one of these levels or, by contrast, it may undergo *mixed lexicalization* if it is affected by various levels simultaneously, e.g. (2e), where lexicalization arises because the *-th* suffix is unavailable but also because the roots in these forms are unproductive (Bauer 1983: 61). The implicit idea is, in short, that there exist degrees in the application of lexicalization and, for this reason, some words are *more* lexicalized than others.

One typical attribute of lexicalized words is opacity, and occurs when lexicalization has applied so intensely that it is impossible to distinguish which elements originally constituted the original formation. Such is the case of *husband*, cited in the *Oxford English Dictionary* (Simpson 2002; hereafter *OED*) as coming from Old English compounding between *hús* ('house') and ?*bónda*, *bonda*, *bunda* ('peasant owning his own house' and 'land, freeholder, franklin, yeoman'). No native speaker of English can guess, by mere observation, that *husband* was formerly a compound, so the lexeme is said to be opaque regarding synchronic morphology. Nonetheless, it has been pointed out (Bauer 1983: 49; Riehemann 1998: 51) that opacity is not a must for lexicalization and, even if both often go hand in hand, there are lexicalized words that are perfectly

analysable morphologically. Examples of lexicalized but transparent words are *warmth* and *width*, both carrying the now unproductive suffix *-th*, as it is possible to split those words by separating their bases (*warm*, *wide*) from the suffix (*-th*). This is so despite the fact that the series of *-th* derivatives cannot be further expanded today (Bauer 1983: 56; Kastovsky 1986: 588).

A preliminary note must be made concerning frequency of usage, believed to be a direct cause of lexicalization (Aronoff and Anshen 2001: 240; Lipka 2002: 111; Bakken 2006: 107). It has been explained that coinages are produced to fulfil a need and may either disappear or become institutionalized and perhaps lexicalized. The assumption underlies this belief that frequency commonly accompanies a word during its progress from one stage onto another, so that coinages will tend to be infrequent, institutionalized words will be more frequent and lexicalized units even more frequent, because they have been around for a longer time.[8] The link between lexicalization and frequency is not absolute and must not necessarily hold, however, and this is the main remark of this paper, groups of lexicalized items have a tendency to carry greater frequencies than synchronically analysable units (see Bauer 2004: 13).

We may associate this idea to the degree of usefulness of units: if speakers do not commonly call for a word, it will be seldom employed, whereas repeated usage will imply greater importance. If so, a unit becomes lexicalized because it has been useful to speakers for longer time than a coinage, hence frequency can be implicitly linked to the importance of a word in time.

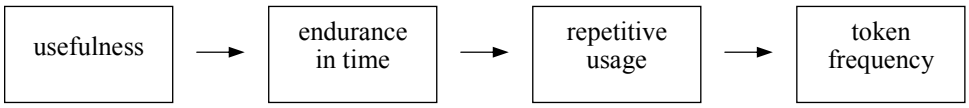

Figure 3. Factors affecting token frequency.

The succession of the factors in Figure 3 brings to light how token frequency and lexicalization correlate, and hints that the ultimate cause of token frequency is usefulness. This does not mean that coinages are less relevant for language users than institutionalized or lexicalized lexemes *at a given time*, but that the latter have been needed *for a longer time*. Because speakers often require having a word for 'a domesticated carnivorous mammal related to the foxes and wolves and raised in a wide variety of breeds', the word *dog* is institutionalized today and is unlikely to disappear in the near future. The opposite happens with a potential verb-forming process with the meaning "grasp

[8] Note that the relationship between the age of a word and its frequency is here put in terms of likeliness, and that there may be individual items where the above does not apply, e.g. *goliard*, which is both rare and old, and has a low token frequency in the *BNC Sampler* (3).

NOUN in the left hand and shake vigorously while standing on the right foot in a 2.5 gallon galvanized pail of corn-meal-mush" (Rose 1973: 516). The perception of the intended meaning is so superfluous that no verb has been coined yet with this semantic configuration, which actually seems improbable to happen sometime soon.

As shown above, certain units are more under the influence of lexicalization than others, so that, for instance, some are morphologically analysable, while others are completely opaque. This gradual nature of lexicalization corresponds with the theoretical and practical schemes of this phenomenon and, as the experiments in section 4 will corroborate, this supports the fact that word-formation processes display different degrees of lexicalization according to the formula π.

Considering the discussion in Sections 2.1 and 2.2, it is hardly surprising that productivity and lexicalization stand as contenders for the achievement of the community's naming needs. If lexicalization has been linked to repetitive usage and endurance in time, productivity symbolizes the dynamicity of language through spontaneous creations, motivated by the necessity of designating some reality. Both notions complement each other when the uttering of a concept is required, and so much so that it has even been argued that morphology and the lexicon are rivals because both serve the same function: they provide words (see Plag 1999: 93; Aronoff and Anshen 2001: 238–242; Dressler 2007: 167).

3. Data preparation

The entries considered for this study belong to the *BNC Sampler*, from where both the lemmas and their frequency have been extracted. As defined in Section 1, this experiment considers exclusively units of the word-class noun, and tests the proposed hypotheses against two sets of items, each believed to be representative of a certain route.

The first group (hereafter subcorpus-ς) is formed by Noun + Noun (hereafter N+N) compounds, a process that embodies usage of the word-formation route. This process seems a suitable candidate for illustration of this route because, due to its very high productivity (Levi 1978: 12; Copestake 2001: 3; Lipka 2002: 95; Lieber 2009: 362), the sample is assumed to contain a high proportion of productively-coined units which, as has been explained above, are generated through the word-formation route. The high productivity of N+N compounding is also the reason to presume that subcorpus-ς will contain a low number of lexicalized words, thus making it representative of productive word-formation.

To compile subcorpus-ς, *Oxford WordSmith Tools* version 4.0.0.376 (Scott 2004) was used in two of its subfunctions. Specifically, the *WordList* and *Concord* applications proved essential at diverse points because, while solid and hyphenated compounds in the *BNC Sampler* carry parts-of-speech (hereafter POS) tagging, open units do not. This entails that they are conceived as syntactic units, which is deduced from the fact that the *BNC Sampler* provides all lexical units with a tag: hence, those N+N constructions

without one must be syntactic objects. For instance, entries such as *airfield*, *birthday* or *carriageway* are identified as NN in the corpus tagging, which automatically sets them as lexical items, in this case noun compounds.

By contrast, in units like *carol concert*, *girls magazine* or *brain damage*, the fact that no hyphen or connector attaches the members of the compound makes the *BNC Sampler* mark each member of the pair separately, and this implies that each noun is seen as a lexical item by itself. Consequently, in cases like these, we are faced not with N+N lexemes but with syntactic constructions where a noun premodifies another noun. This means that querying the corpus for all compound nouns with *WordList* will not return N+N units like *carol concert* because they lack POS-tagging, a major drawback to this tagging.

These features of N+N units raise the question of the morphology-syntax interface, one of the most complex issues in current linguistics, and one where the number of publications is equalled by the disagreement of the linguists concerned with it (see Bauer 1998; Plag et al. 2006; Giegerich 2009; Lieber and Štekauer 2009). The literature has been engaged for years in discussing whether N+N units fall under the domains of morphology or of syntax and, what is even more complex, in detecting the boundaries between both modules in a practical manner. Despite the lack of consensus over this topic, one conclusion from the specialists is that N+N units are perceived as having a morphological or a syntactic nature depending on the angle we take, as one may focus on their phonology, orthography, semantics or morphology (see Fernández-Domínguez 2009: 27–41). Due to the varying results which such tests provide, it seems that a joint combination of these factors should be considered before attempting in-depth analyses in this sense. Bauer (1998: 78) puts it as follows:

> [A]ny distinction drawn on the basis of just one of these criteria is simply a random division of noun+noun constructions, not a strongly motivated borderline between syntax and the lexicon.

Bearing all the above in mind, *WordList* was used together with the POS-tagging for the retrieval of solid and hyphenated compounds, while *Concord* was employed for open N+N units. Such a method is beneficial because the resulting preliminary list comprises all potentially relevant N+N compounds (about 6,000 entries), something not viable otherwise.

It is also true, however, that *Concord* returns all open N+N combinations regardless of the context of the unit or the nature of its constituents and, as a result, many of these constructions were not really compounds, but items with no relevance for this study. For this reason, filters were applied to discard entries others than N+N compounds, which includes the following:

(i) Formations where an adjective premodifies only the first constituent of the compound, like *[low] alcohol wine*, *[old] age pensioner*, *[open] book exam*.

(ii) Units including proper nouns, like *ITT Avionics Division* or *Premier Cup*.
(iii) Exocentric compounds, like *cheese factor* and *couch potato*.
(iv) Synthetic compounds, like *air freshener* or *arms-trading*.

Attention was also paid here to preserve exclusively entries with a degree of lexicalization sufficient to be regarded as lexemes, like *beer belly*, *bomb threat*, *gas layer*, *peace process* or *spy satellite*, and not just any random concatenation of two nouns (*arch rivals*, *chin chimney*, *clip clop*). In addition, the context of occurrence of units was checked in cases of ambiguity to separate noun phrases from N+N compounds and, when any change was made, the frequency figures were adapted accordingly.

These filters were designed to keep subcorpus-ς as uniform and homogenous as possible, while making it fully compatible with the requirements of Levi's (1978) *Recoverably Deletable Predicates* (hereafter, RDPs), which are suitable exclusively for regular and non-lexicalized N+N compounds (see fn. 9). In the case of lexicalized nouns, they were set apart but not deleted, as they were later useful for the compilation of the second subcorpus (see below in this section). After application of these filters, some 900 entries are discarded, after which subcorpus-ς encompasses roughly 5,000 N+N compounds.

As regards lexicalized units (hereafter, subcorpus-λ), they were chosen because in principle they are generated as unproductive output, and should thus be representative of the lexicon route (see Section 2.2). The objective was to obtain units unanimously agreed to have undergone lexicalization, regardless of whether different word-formation rules were originally involved in their creation. The reason why units from different processes are relevant is that the lexicon route is employed for the uttering of any lexicalized unit, no matter which process created it. Consequently, the fact that a unit is a compound, a converted lexeme or an affixed item is not significant for these purposes: all of them can be equally affected by frequency and, therefore, by lexicalization too (see Bauer 1983: 61).

Considering the lack of well-defined lines to separate lexicalized from institutionalized items (see Sauer 2004: 1627), the retrieval of entries was done based on specialized publications: Bauer (1983), Lipka (2002), Sauer (2004), Brinton and Traugott (2005) and Hohenhaus (2005). In this case, the lexicalized units put aside during the creation of subcorpus-ς were picked up and checked against these publications. A lexicalized unit was incorporated into subcorpus-λ so long as

(i) all publications agreed on its lexicalized status, or
(ii) at least two publications considered it a lexicalized unit and no publication rejected it.

Point (i) can be illustrated with *-th* derivatives (*warmth*, *length*, *width*), which are unanimously regarded as lexicalized by the above-mentioned references and stand as classical examples of this phenomenon; accordingly, these items were added to subcor-

pus-λ. By contrast, a unit like *trouserpocket* is quoted as lexicalized only by Hohenhaus (2005: 356) and, because no other source mentions it, it was discarded for this analysis. This is so even if *trouserpocket* is a clear case of lexicalization because the otherwise obligatory plural morpheme *-s* has been dropped from the base word *trousers*.

As for point (ii), we may give the example of *blackbird*, cited by Bauer (1983) and Brinton and Traugott (2005) as a case of lexicalization. In this case, two sources back up the status of the unit and, because no other source conflicts with that view, *blackbird* is allowed into subcorpus-λ.

When all lexicalized entries were retrieved, filters were applied not to include doubtful or borderline cases, but exclusively widely-accepted examples of lexicalization. This stage discards:

(i) units from word-classes others than nouns, for example adjectives (*durable*, *aggressive* or *infamous*);
(ii) invented words (*Kodak* or *quark*); and
(iii) records which are not actual English words (*£1*, ***base/basis* or *50%*).

All units ready, the *BNC Sampler* was used to track down each corpus entry plus its individual frequency, which involves the consideration not only of the citation forms of lemmas (i.e. singular), but also of the diverse realizations in inflection and spelling. The inspection of spelling variants affects both subcorpus-ς and -λ, and examines the alternatives singular/plural and solid/hyphenated/open in both members of the compound (solid singular, solid plural, hyphenated singular, hyphenated plural, etc.), which means making an average of 5 *BNC Sampler* searches per entry. The frequencies of each variant are summed up for the definitive entry and, once all variants are inspected, the formal realization retained is that of the unit with the highest frequency. In (3), (f) represents the most common formal realization and features the addition of all partial frequencies:

(3a)	wheelchair	602
(3b)	wheelchairs	178
(3c)	wheel-chair	10
(3d)	wheel chair	4
(3e)	wheel chairs	3
(3f)	WHEELCHAIR	797

It is at this point that the previous use of *Concord* proves valuable because the list compiled then now allows considering open realizations of solid/hyphenated compounds, thus providing a comprehensive analysis of the corpus figures through all possible realizations of a given lexeme.

Once the frequency of all lemmas is retrieved, subcorpus-λ is available with approximately 870 lexicalized units. As is well-known (Bauer 1983: 52; Štekauer 2000:

81–82; Sauer 2004: 1628), lexicalization very often affects compounds, and this is why most corpus records (65%) were originally created by this process. There are, for example, primary compounds with the configuration Adjective + Noun (4a) (hereafter, AJ+N), Noun + Noun (4b), or Verb + Noun (4c) (hereafter, V+N), as well as synthetic compounds 0(5). Nevertheless, affixation is also present in the corpus, especially in lexemes derived by *-ment* (6a) and *-th* (6b):

(4) (a) blackboard
sweetmeat

(b) country house
cupboard

(c) breakfast
pushchair

(5) shortcoming
streetwalker

(6) (a) achievement
amusement

(b) warmth
width

After following all pertinent procedures, 5,878 lexemes are available for the testing of our hypotheses: 5,010 forming subcorpus-ς, 868 forming subcorpus-λ. In keeping with the theoretical assumptions in Section 2, and based on the aforementioned methodological stages, it seems safe to assert that the 5,010 units in subcorpus-ς characterize usage of the word-formation route, while the 868 entries in subcorpus-λ embody the lexicon route (see Figure 1). This belief derives from the fact that subcorpus-ς contains N+N compounds, units which, as the literature points out (e.g. Lieber 2009: 362), are one of the most fertile lexical devices of Contemporary English. If so, and considering the representativeness of a corpus like the *BNC Sampler*, it follows that subcorpus-ς must display a predominance of synchronically-analysable entries over lexicalized items. This naturally leads to perceiving subcorpus-ς as illustrative of the word-formation route.

Subcorpus-λ, in a complementary manner, is made up of lexicalized units, i.e. those which have lost their syntagmatic nature and have to be listed in the lexicon for understanding (see Riehemann 1998: 51; Lipka 2002: 97; Bauer 2004: 12–18). These units, thus, are taken to be uttered through the lexicon route, given that the specialized publications quoted above judge that they cannot be analyzed following the principles of synchronic word-formation. It is essential to understand that the lexicalized lexemes are a part of this corpus not necessarily because they belong to currently unavailable processes, but because they themselves are lexicalized. For instance, the units derived by *-ment* (*argument*, *employment* or *involvement*) are lexicalized insofar as this process is unproductive in Contemporary English. On the other hand, units like *blackboard*, *holiday* or *mincemeat* are lexicalized because their internal semantics are irregular and unpredictable, even if AJ+N compounding is able to coin new items today (see Bauer 1983: 61; Lipka 2002: 113; Sauer 2004: 1632).

The last methodological step performs a semantic analysis using Levi's (1978) RDPs. Her nine-category set is operated on subcorpus-ς and allows discerning the meaning relationship between the two components of N+N compounds, something otherwise impossible given the structural makeup of these units (see Levi 1978: 5–12; Bauer 1983: 159–163; Štekauer 2000: 58–59; Jackendoff 2002: 249–250). The following illustrates the nine RDPs using examples from the study corpus, as well as the membership distribution of each of them.[9]

Table 1. Illustration of Levi's (1978) RDPs.

RDPs	Example	Paraphrase	Units
ABOUT	*accident report*	the report is *about* the accident	787
BE	*ozone layer*	the layer *is* ozone	508
CAUSE	*tomato disease*	the disease is *caused* by tomatoes	183
FOR	*mousetrap*	the trap is *for* mice	955
FROM	*bearskin*	the skin comes *from* the bear	159
HAVE	*beefburger*	the burger *has* beef	562
IN	*Easter egg*	the egg is eaten/made *at* Easter	1181
MAKE	*silkworm*	the worm *makes* silk	371
USE	*card game*	the game *uses* cards	304
		Total	5010

This approach allows allocating each corpus entry to one RDP depending on the entry's individual load of meaning, which in turn makes it possible to construct clusters of similar items after the inspection of all corpus records. During this process, there were admittedly specific lexemes where more than one semantic interpretation was viable, therefore special care was taken for the definitive distribution of lexemes across RDPs given the magnitude of this semantic breakdown. A case of semantic vagueness is shown here for the compounds *church magazine* (7) and *door mat* (8), which have two possible readings depending on the context:

(7a) 'the magazine is *in* the church' RDP IN
(7b) 'the magazine is *about* the church' RDP ABOUT

[9] Lexicalized units are excluded from this analysis because Levi (1978: 8) stresses that her semantic analysis is aimed at "nonlexicalized, nonspecialized, nonidiomatic, and [...] nonmetaphorical forms", which explicitly reduces the scope of the RDPs and makes lexicalized units inappropriate at this point.

(8a) 'the mat is *for* the door' RDP FOR
(8b) 'the mat is *at* the door' RDP IN

In cases like these, the *BNC Sampler* was used to track down the contexts of the entries in question, a procedure that made it possible to remove semantic ambiguity and then assign a given RDP to each entry. This stage also considered the possibility of having polysemous lexemes, for which the contexts in the *BNC Sampler* were also used. Still, no instance of polysemy was found in the study corpus, hence no further analysis was required in this sense.

After completion of the RDP-based semantic analysis, we can refer to FROM N+N compounds or to HAVE N+N compounds globally, hence embracing all units with analogue meanings. This seems to be, at present, the only viable alternative to perform a semantic analysis on N+N compounds because, unlike in affixation, there is no detachable morpheme to disclose the new formation's structure. While affixes like *-al*, *-ity* or *-ness* are by themselves formal signs that word-formation has applied (e.g. *great* + *-ness* = *greatness*), there is no mark in *cat muck* or *furniture shop* to reveal that a new word has been created. Use of the RDPs, then, is fundamental for productivity computations of N+N compounds because, despite their widely-agreed high productivity, no study has tried to compute their productive potential yet, to the best of my knowledge.

Note that, in spite of these advantages, it is futile to encompass all the existing meaning relationships of N+N compounds by use of lists like Lees (1960), Li (1971) or Levi (1978), according to the results in Downing (1977). Based on the reading of novel compounds by informants, Downing shows that such sets of fixed predicates are valid if they are regarded to only underlie the meanings of new units, as they are reduced versions of the actual meanings of compounds and therefore imply "the loss of much of the semantic material" (Downing 1977: 826). Downing's work was an early pioneer not only in the study of the semantics of N+N compounds, but also in their interpretation both in and out of context, a subject that has been recently taken on in monographs like Štekauer (2005) or Kuperman et al. (2008).

In this case, however, Levi's (1978) analysis proves essential for the application of π because this formula needs to be used on analogue sets of items. Hardly any conclusion could be drawn from an all-encompassing operation on the entire corpus, since an isolated figure would not be significant in semantic terms. In Section 4, we argue for a distinct origin of each of these groups and how their frequencies expose the route that they have taken.

4. Evidence for usage of the word-formation vs. the lexicon route

As put forward in Section 2.1, a naming need may be fulfilled either by use of the lexicon or by coinage of a new item through word-formation, and each alternative has different implications for the frequencies of a process. The argument here is that use of the

above two routes is echoed by corpus frequencies because these are directly linked to *V* and *N* in a representative corpus.

A premise sketched in Figure 1 is that the lexicon route involves the repetitive use of a lexeme, whereas the word-formation route is an only-once act which implies the creation of a new unit. It can be argued, with such a scheme in mind that, whenever a unit is generated through the word-formation route, this will mean an increase of 1 in the type frequency of that particular rule, as one new unit has been created which now counts as a type towards that process (see Section 2.1). Then, considering the features of both routes, it can be realized that the mean token frequency is higher in non-productive processes than in productive processes because in the former the same words are used constantly, whereas the latter produce new words and frequency is prone to disperse (see Anshen and Aronoff 1981; Romaine 1983: 180; Baayen and Lieber 1991: 830; Bauer 2001: 152; Namer 2003: 80).

This can be demonstrated by combining type and token frequency under the belief that type frequency will tend to be more favourable to profitable word-formation processes, while in unproductive processes the situation will be the opposite (Aronoff 1983). Put differently, the more types are coined the lower the distribution of token frequency is among them. In contrast, when a process is less profitable, fewer individual units are created (lower *V* value), the same derivatives being used constantly (higher *N* value). What follows is that the type/token ratio tends to bolster types in productive processes and tokens in less productive processes, which can be formalized through the indicator of profitability (π):

$$\pi = \frac{V}{N}$$

where the higher the final figure the higher the profitability of the process in question, and vice versa. A high figure is found where the ratio between *V* and *N* is favourable to *V*, which suggests that more units from the process in question have taken the word-formation route. By the same token, the lower the result the higher *N*, with the implication that more units from that process use the lexicon route. The highest possible value in this formula is 1, and happens when one type has been used as infrequently as possible, i.e. once ($N = 1$). Note that *V* will never be higher than *N*, as there cannot be more different units than uses of those units.

We may now approach the study sample to confirm the foreseen guesses. Let us first focus on subcorpus-λ. Here, six major groups of units can be distinguished depending on their original word-formation process: *-ment*, *-th*, AJ+N compounds, N+N compounds, V+N compounds and synthetic compounds.

Table 2 shows the results of π for these six groups.

As is manifest, these figures remarkably match the facts as regards the status of lexicalization of each process. Three out of the six processes examined are unavailable in Contemporary English (V+N, *-ment* and *-th*; see Bauer 1983: 56, 2001: 205), one is

Table 2. Results for π in subcorpus-λ.[10]

WFP	V	N	π
Synthetic	94	45,116	0.002083
N+N	397	549,511	0.000722
AJ+N	132	214,225	0.000616
V+N	189	322,436	0.000586
-ment	53	91,379	0.000580
-th	3	34,210	0.000087

considered to be marginally productive (AJ+N; see Giegerich 2005: 3, Lieber 2009: 362) and the other two are very profitable (N+N and synthetic compounding; see Lieber 2009: 360–361), and this is precisely the ranking provided by π. Correspondingly, synthetic compounding leads the ranking with the highest value (0.002083), followed by N+N compounding (0.000722) and AJ+N compounding (0.000616). These three processes are available nowadays, and are followed in order by the unavailable ones: V+N compounding (0.000586), *-ment* (0.000580) and *-th* (0.000087).

These results disclose that the three top-ranked processes include items employing the lexicon route but they also contain synchronically-analysable items, while processes like *-th* consist exclusively of lexicalized formations which are obligatorily expressed through the lexicon. This means that a higher share of the set of synthetic compounds than of the set of V+N compounds is used productively at present, precisely because the latter process is unavailable, while the former is still fertile. The above can be attested by looking at the study corpus, as the three top-ranked processes comprise lexemes which, although lexicalized, have low token frequencies, and this is favourable to π: *streetwalker* (13) and *sleepwalker* (26) in synthetic compounding, *halibut* (30) and *sweetmeat* (42) in AJ+N compounding, and *fishwife* (18) and *pyjama top* (9) in N+N compounding.

A significant point about Table 2 is that it is consistent with the belief that lexicalization does not either happen or not, but that there is a *gradual listing* whereby some units may be more lexicalized than others. According to the above results, groups with lower values contain lexemes where lexicalization has applied more strongly, or more lexicalized lexemes in general. By contrast, higher values suggest that lexicalization is not so widespread among the units of a given cluster, or that, when it is, it has not applied intensely. This idea has been expressed, for example, by Jackendoff, who does not find mandatory to have "a strict cutoff in frequency between stored forms and forms generated on-line" but, rather, "a cline in accessibility for stored forms, including the usual differential between recognition (better) and recall (worse)" (Jackendoff 1997: 231; see Bauer 1983: 61; Hohenhaus 2005: 368).

[10] The results are here sorted out from more to less productive according to π. The same applies to subsequent tables.

It is also possible, by looking at the above figures, to detect an approximate value for the borders of availability in this study, i.e. the lowest figure which a process must reach to be available. Given the above results, this border must be located between values 0.000586 and 0.000616, because those are the respective figures of V+N compounding (currently unavailable) and AJ+N compounding (currently available). This can be confirmed by looking at Table 2 because all processes above that line are available, while those below it are not. Availability is an indispensable but also neglected question due to the shortage of qualitative studies in the field of productivity, traditionally assessing which processes are the most productive but not which are on the cusp between productivity and unproductivity. It is then to be expected that contributions to come make up for the lack of attention to availability by considering alternative approaches to this concept.

Also note that π is able to distinguish between attestation (i.e. the testimony of existence of a word) and use of the lexicon route because, although one process may encompass more items than another (i.e. *V*), what really matters is the relationship between types and tokens (Copestake 2001: 2). This is evident in that processes with more types are not necessarily listed higher than processes with fewer types, as in synthetic compounding, which leads the ranking and has only *V* 94. As a token, V+N compounding consists of 189 types but it nevertheless displays a poor value, and this confirms that *V* is useful to measure attestation, while π measures distribution of high-frequency items.

Formula π may be applied to subcorpus-ς as well. In this case, by use of Levi's (1978) RDPs, the 5,010 N+N compounds are divided into the nine semantic categories that allow gauging their profitability. It is fundamental to note that π measures something different here than in Table 2. While there, the formula is used to gauge the degree of lexicalization of processes, the goal of π here is to calculate profitability, as N+N compounding is acknowledged to be available nowadays and all entries in subcorpus-ς come from this process. Thus, although there exist lexicalized N+N compounds (see Table 2), all the entries in this subcorpus are produced by the word-formation route, and their profitability values will expectedly be considerably high, as shown in Table 3.

Table 3. Results for π in subcorpus-ς.

RDP	V	N	π
CAUSE	183	2,921	0.0626
IN	1,181	69,044	0.0171
MAKE	371	22,573	0.0164
FOR	955	59,494	0.0160
HAVE	562	41,853	0.0134
USE	304	23,876	0.0127
FROM	159	13,483	0.0117
ABOUT	787	74,719	0.0105
BE	508	73,820	0.0068

What is noteworthy, first, is the variation between these figures and those of N+N compounding in Table 2, where this process has a value of 0.00072. On this occasion, however, all nine categories greatly surpass that figure, and even the lowest value (BE, 0.0068) exceeds it to a large extent. Bearing in mind the rationale of π, this divergence denotes that the share of lexicalized/institutionalized items is much larger in the units in Table 2 than in Table 3, which leads to a low figure of π in the former and, in turn, to claim that the latter has a higher rate of profitability.

Internal observation of the RDPs, second, brings about a couple of unpredicted facts. For example, that CAUSE (0.0626) is listed on top of the list may be surprising given that it is usually considered a poorly productive predicate (Levi 1978: 87–88). A logical question follows: why does π set it as the most profitable of the set? The answer lies in the figures of *V*. It was explained in Section 2.1 that profitability has a quantitative nature, linked with the number of coined items, and that, although it is different from attestation of forms, there are nuances common to both. To be precise, profitability bears a close relationship with frequency because, although not directly comparable, high profitability leaves traces in the type frequency of processes. This happens when a process generates a sufficient number of coinages and they become widespread in use; logically, many of them become established and embody different types in a corpus (see Neuhaus 1973; Plag 1999: 33–34; Bauer 2001: 48–49, 144–145; Namer 2003: 89; Dressler 2007: 163).

A consequence is that, as π computes profitability (i.e. the quantitative side of productivity), a *Minimum V-Input* (hereafter MVI) is needed for reliable results, otherwise its findings will be distorted. For a given experiment, the MVI is the smallest possible *V* figure which a process needs for π to be operated on it successfully. This general rule imposes the prerequisite of having a high enough *V* value, which is, in fact, a coherent condition because having a low *V* indicates low profitability by itself, therefore any further operation in this sense is unneeded. The MVI can be detected by observing the figures of π, which reveal whether the number of entries of some clusters is insufficient for consistent findings. For example, CAUSE, with 183 as *V* value, triggers π to operate illogically by providing this predicate with a very high figure. The analyst, therefore, must be aware of these facts and adjust the results by placing CAUSE (0.0626) and FROM (0.0117) as the least profitable predicates due to their inadequate MVI. From previous studies (Fernández-Domínguez 2008), the MVI seems to significantly hang on the number of sampled items so that, in general, smaller samples tend to have smaller MVIs. The precise figure ranges between 1.9% and 4.12% of the corpus size;[11] in this case, it is 3.6% (183 units out of 5,010).

The remaining clusters in Table 3 occupy more coherent positions, the most profitable being IN (0.0171), then MAKE (0.0164), FOR (0.0160), HAVE (0.0134), USE (0.0127), ABOUT (0.0105) and BE (0.0068). An aspect of this inspection is the gradual

[11] The MVI was determined at 60 for a corpus of 3,093 entries (i.e. 1.9%), while it was set at 71 in a corpus of 1,720 items (i.e. 4.12%).

progression in the predicates' figures, such that the differences between entries are minor, especially in the range between IN (0.0171) and ABOUT (0.0105), where predicates present very close areas of profitability and only 0.0066 separates the most profitable cluster from the least profitable one. This is relevant in that seven out of the nine predicates appear within these boundaries, which means that their profitability, although ranked differently, is relatively similar. In her work, Levi (1978: 77–80) writes that the most productive RDPs are FOR and IN, while CAUSE and FROM stand at the other end of the line, facts which can be corroborated through the figures in Table 3.

The adequacy of π can also be exposed by observing predicates MAKE and FOR, whose values indicate a practically equivalent profitability despite their different *V* figures (371 in MAKE, 955 in FOR). This confirms π's awareness that attestation does not necessarily lead to profitability (see Plag 1999: 25), and so word-formation processes with a high type frequency may nevertheless be seen as poorly profitable.

In view of the above discoveries, two prerequisites seem to be necessary for successful application of π in synchronic processes:

(i) that the process studied is available, i.e. processes with the feature [–AVAILABLE] do not fall under the scope of this formula, hence their results may be erroneous in this sense,
(ii) that the process has a sufficient number of *V*, so that the type/token ratio can be reliably operated and the distribution of items across the corpus is a uniform one. The MVI is dependent on the corpus size, and can be set at around 4.12% from previous experiments. Where this value is deficient, this is by itself indicative of the low profitability of the process in question (see cause in Table 3).

5. Conclusions

Lexicalization and productivity are intimately related to type and token frequency, a connection which has been habitually surveyed in the literature (Bauer 1983: 48–49, 2001: 47–49; Plag 1999: 24–34; Aronoff and Anshen 2001: 240; Lipka 2002: 111). Despite how often word-formation studies have demonized lexicalization because of its alleged irregularities, units stored in the lexicon offer valuable data that can be contrasted with regularly productive patterns in different ways. It has been shown, for example, that some lexicalized units are morphologically regular, and these can perform a task in anticipating the form of future coinages if we believe that it is possible for listed entries to influence how new words are created. The significance of lexicalization is remarked by Riehemann (1998: 51–52; see Verspoor 1998: 161; Allan 2006: 148): "Approaches that do not relate lexicalized words to rules in any way also do not predict that existing words in the language will have an effect on what the rules of morphology are and do not have anything to say about how they could be acquired".

This paper has argued for a strong link between corpus-based frequencies and the use of two possible routes for the output of naming needs. In particular, two experiments have been conducted for the attainment of the objectives in 1: the first looks at the correlation between lexicalization and frequency and, by focusing on subcorpus-λ, provides an approximate ranking of lexicon use. The point under consideration, namely that use of the lexicon route is typical of unproductive processes, is validated through π by correctly positioning various word-formation processes according to their degree of institutionalized/lexicalized units.

The second experiment uses subcorpus-ς to measure profitability. The approach taken here is to semantically analyse these entries through Levi's (1978) RDPs and then to operate π on them. Again, π is used to assess the peaks of profitability in the predicates, with results that lead to certain conclusions:

(i) The results of π refer to profitability, not to productivity in a global sense. The reason is that morphological productivity is stressed here to be an epiphenomenon, i.e. "[…] a property that results from other mechanisms" (Plag 1999: 12), which emphasizes the bipolar nature of productivity (availability and profitability). This explains why the availability of the process at issue must be studied before measuring its profitability, as it is impossible to grasp both subconcepts with only one formula.

(ii) This model puts in a fine performance in relation to the distinction profitability vs. attestation, as proved by the fact that having a high *V* value does not ultimately determine the figure of π. What can be highlighted here, rather, is that π is able to correctly calculate the average of units that use the word-formation route for a process (see in vs. make in Table 3). This is important in that π is more than a straight frequency-productivity correlation, which scholars have often refused (see Plag 1999: 111–115, 2006: 550; Bauer 2001: 48–49; Štekauer 2005: 88–90).

(iii) The study of word-formation seems to be more accurately approached from a semantic (or, at least, formal-semantic) point of view than from a strictly formal one. Past studies (Štekauer 2001; Fernández-Domínguez et al. 2007) have provided evidence that formal structure does not always match semantic content, and this requires taking certain precautions in the analysis of, for example, affixation. Along these lines, a semantic analysis, Levi (1978), was applied before attempting any productivity calculation. This guarantees that the results obtained are relevant for homogeneous semantic groups, which allows measuring productivity in a consistent manner (see similarly Štekauer 2005: 46–51).

(iv) Frequency calculations of the present type can provide a picture of past or present productivity, but it is doubtful that they can predict future productivity straightforwardly (see Plag 2006: 540–541). One may guess that a present tendency of productivity will continue in the future and also predict the well-formedness of a potential coinage, but it is utterly impossible to foresee which processes will coin more words and which will coin fewer words. This belief emerges from a view of

productivity as emanating from the speakers' naming needs: if naming needs are unpredictable, so should morphological productivity be.

REFERENCES

Allan, K. 2006. "Lexicon: Structure". In: Brown, K. (ed.), *Encyclopedia of language and linguistics* (2nd ed.). Oxford: Elsevier. Vol. 7, 148–151.

Anshen, F. and M. Aronoff. 1981. "Morphological productivity and phonological transparency". *Canadian Journal of Linguistics* 26. 63–72.

Aronoff, M. 1976. *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Aronoff, M. 1983. "Potential words, actual words, productivity and frequency". *Proceedings of the 13th International Congress of Linguists*, Tokyo. 163–171.

Aronoff, M. and F. Anshen. 2001. "Morphology and the lexicon: Lexicalization and productivity". In: Spencer, A. and A.M. Zwicky (eds.), *The handbook of morphology* (Blackwell Handbooks in Linguistics). Oxford: Basil Blackwell. 237–247.

Baayen, R.H. 2005. "Morphological productivity". In: Köhler, R., G. Altmann and R.G. Piotrowski (eds.), *Quantitative linguistics: An international handbook*. Berlin: Mouton de Gruyter. 243–255.

Baayen, R.H. and R. Lieber. 1991. "Productivity and English derivation: A corpus-based study". *Linguistics* 29. 801–844.

Bakken, K. 2006. "Lexicalization". In: Brown, K. (ed.), *Encyclopedia of language and linguistics* (2nd ed.). Oxford: Elsevier. Vol. 7, 106–108.

Bauer, L. 1983. *English word-formation*. Cambridge: Cambridge University Press.

Bauer, L. 1998. "When is a sequence of two nouns a compound in English?" *English Language and Linguistics* 2. 65–86.

Bauer, L. 2001. *Morphological productivity*. Cambridge: Cambridge University Press.

Bauer, L. 2004. "Adjectives, compounds and words". *Nordic Journal of English Studies* 3(1). 7–22.

Bauer, L. 2005. "Productivity: Theories". In: Štekauer, P. and R. Lieber (eds.), *Handbook of word-formation* (Studies in Natural Language and Linguistic Theory 64). Dordrecht: Springer. 315–334.

Bauer, L. and R. Huddleston. 2002. "Lexical word-formation". In: Huddleston, R. and G.K. Pullum (eds.), *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press. 1621–1721.

Bloomfield, L. 1933. *Language*. Chicago: University of Chicago Press.

Brinton, L.J. and E.C. Traugott. 2005. *Lexicalization and language change*. Cambridge: Cambridge University Press.

Carstairs-McCarthy, A. 1992. *Current morphology*. London: Routledge.

Chomsky, N. 1964. *Current issues in linguistic theory*. Den Haag: Mouton de Gruyter.

Chomsky, N. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Copestake, A. 2001. "The semi-generative lexicon: Limits on lexical productivity". In: Bouillon, P. and K. Kanzaki (eds.), *Proceedings of the First International Workshop on Generative Approaches to the Lexicon*, 15–24.
<http://www.cl.cam.ac.uk/~aac10/ papers/glex2001.pdf> (Last accessed 12 November 2009.)

Corbin, D. 1987. *Morphologie dérivationelle et structuration du lexique*. Tübingen: Max Niemeyer.

Dal, G. 2003. "Productivité morphologique: Définitions et notions connexes". *Lange Française* 140. 3–23.
De Jong, N.H., R. Schreuder and R.H. Baayen. 2000. "The morphological family size effect and morphology". *Language and Cognitive Processes* 15. 329–365.
Di Sciullo, A.M. and E. Williams. 1987. *On the definition of word*. Cambridge, MA: MIT Press.
Dokulil, M. 1962. *Tvoření slov v češtině 1. Teorie odvozování slov*. Praha: ČAV Praha.
Downing, P. 1977. "On the creation and use of English compound nouns". *Language* 53(4). 810–842.
Dressler, W.U. 2003. "Degrees of grammatical productivity in inflectional morphology". *Rivista di Lingüística* 15(1). 31–62.
Dressler, W.U. 2007. "Productivity in word-formation". In: Jarema, G. and G. Libben (eds.), *The mental lexicon. Core perspectives*. Amsterdam: Elsevier. 159–183.
Fernández-Domínguez, J. 2008. Productivity measurement of English compounding based on a corpus of the nominal type. (Unpublished PhD dissertation, University of Jaén.)
Fernández-Domínguez, J. 2009. *Productivity in word-formation. An approach to N+N compounding*. Bern: Peter Lang.
Fernández-Domínguez, J., A. Díaz-Negrillo and P. Štekauer. 2007. "How is low morphological productivity measured?" *Atlantis* 29. 29–54.
Giegerich, H.J. 2005. "Lexicalism and modular overlap in English". *SKASE Journal of Theoretical Linguistics* 2(2). 571–591.
<http://www.pulib.sk/skase/Volumes/JTL03/ 06.pdf> (Last accessed 3 September 2009.)
Giegerich, H.J. 2009. "The English compound stress myth".
<http://www.englang.ed.ac.uk/ people/heinz.html> (Last accessed 26 July 2009.)
Hay, J. 2003. *Causes and consequences of word structure*. London: Routledge.
Hay, J. and R.H. Baayen. 2005. "Shifting paradigms: Gradient structure in morphology". *Trends in Cognitive Sciences* 9(7). 342–348.
Hohenhaus, P. 2005. "Lexicalization and institutionalization". In: Štekauer, P. and R. Lieber (eds.), *Handbook of word-formation* (Studies in Natural Language and Linguistic Theory 64). Dordrecht: Springer. 353–373.
Jackendoff, R. 1997. *The architecture of the language faculty*. Cambridge, MA: MIT Press.
Jackendoff, R. 2002. *Foundations of language. Brain, meaning, grammar, evolution*. Oxford: Oxford University Press.
Kastovsky, D. 1986. "The problem of productivity in word-formation". *Linguistics* 24. 585–600.
Kuperman, V., R. Bertram and R.H. Baayen. 2008. "Morphological dynamics in compound processing". *Language and Cognitive Processes* 23. 1089–1132.
Lees, R.B. 1960. *The grammar of English nominalizations*. Bloomington, IN: Indiana University Press.
Levi, J.N. 1978. *The syntax and semantics of complex nominals*. New York: Academic Press.
Li, C. 1971. Semantics and the structure of compounds in Chinese. (Unpublished PhD dissertation, University of California, Berkeley.)
Lieber, R. 2009. "IE, Germanic: English". In: Lieber, R. and P. Štekauer (eds.), *The Oxford handbook of compounding* (Oxford Handbooks in Linguistics). Oxford: Oxford University Press. 357–369.
Lieber, R. and P. Štekauer. 2009. "Introduction: Status and definition of compounding". In: Lieber, R. and P. Štekauer (eds.), *The Oxford handbook of compounding* (Oxford Handbooks in Linguistics). Oxford: Oxford University Press. 3–18.
Lipka, L. 1977. "Lexikalisierung, Idiomatisierung und Hypostasierung als Probleme einer synchronischen Wortbildungslehre". In: Brekle, H.E. and D. Kastovsky (eds.), *Perspektiven der Wortbildungsforschung*. Bonn: Bouvier. 155–164.

Lipka, L. 2002. *English lexicology: Lexical structure, word semantics and word-formation*. Tübingen: Gunter Narr.
Marcus, G., M. Ullman, S. Pinker, M. Hollander, T.J. Rosen and F. Xu. 1992. *Overregularization in language acquisition.* (Monographs of the Society for Research in Child Development 57.)
Namer, F. 2003. "Productivité morphologique, représentativité et complexité de la base: Le système MoQuête". *Langue Française* 140. 79–101.
Neuhaus, H. J. 1973. "Zur Theorie der Produktivität von Wortbildungssystemen". In: Ten Cate, A.P. and P. Jordens (eds.), *Linguistische Perspektiven: Referate des VII Linguistischen Kolloquiums*. Tübingen: Max Niemeyer. 305–317.
Nishiwaka, M. 2003. "Lexicon and cognition. A study of listed syntactic objects". *The Humanities* 52. 29–39.
Pinker, S. 1991. "Rules of language". *Science* 253. 530–535.
Pinker, S. and A. Prince. 1988. "On language and connectionism: Analysis of a parallel distributed processing model of language acquisition". *Cognition* 28. 73–193.
Plag, I. 1999. *Morphological productivity: Structural constraints in English derivation.* Berlin: Mouton de Gruyter.
Plag, I. 2003. *Word-formation in English*. Cambridge: Cambridge University Press.
Plag, I. 2006. "Productivity". In: Aarts, B. and A. McMahon (eds.), *Handbook of English linguistics* (Blackwell Handbooks in Linguistics). Oxford: Basil Blackwell. 537–556.
Plag, I. and R.H. Baayen. 2009. "Suffix ordering and morphological processing". *Language* 85. 109–152.
Plag, I., G. Kunter, S. Arndt-Lappe and M. Braun. 2006. "Modeling compound stress in English".
<http://www2.uni-siegen.de/~engspra/DFG-Project/abstract.htm> (Last accessed 4 November 2008.)
Riehemann, S.Z. 1998. "Type-based derivational morphology". *Journal of Comparative Germanic Linguistics* 2. 49–77.
Romaine, S. 1983. "On the productivity of word formation rules and limits of variability in the lexicon". *Australian Journal of Linguistics* 3. 177–200.
Rose, J.H. 1973. "Principled limitations on productivity in denominal verbs". *Foundations of Language* 10. 509–526.
Rumelhart, D.E. and J.L. McClelland. 1987. "Learning the past tenses of English verbs: Implicit rules of parallel distributed processing". In: MacWhinney, B. (ed.), *Mechanisms of language acquisition*. Mahwah, NJ: Erlbaum. 194–248.
Sauer, H. 2004. "Lexicalization and demotivation". In: Booij, G.E., C. Lehmann, J. Mudgan and S. Skopeteas (eds.), *Morphologie. Morphology. Ein internationales Handbuch zur Flexion und Wortbildung. An international handbook on inflection and word-formation. 2. Halbband.* (Volume 2.) Berlin: Walter de Gruyter. 1625–1636.
Schultink, H. 1961. "Produktiviteit als morfologisch fenomeen". *Forum der Letteren* 2. 110–125.
Scott, M. 2004. *Oxford WordSmith Tools*, version 4.0.0.376. Oxford: Oxford University Press.
<http://www.lexically.net/wordsmith/version4/index.htm> (Last accessed 10 May 2010.)
Selkirk, E.O. 1982. *The syntax of words*. Cambridge: MIT Press.
Simpson, J.A. (ed.). 2002. *The Oxford English Dictionary on CD-ROM*, version 3.1 (2nd ed.). Oxford: Oxford University Press.
Spencer, A. 1991. *Morphological theory: An introduction to word structure in generative grammar*. Oxford: Basil Blackwell.
Štekauer, P. 2000. *English word-formation. A history of research (1960–1995)*. Tübingen: Gunter Narr.

Štekauer, P. 2001. "Fundamental principles of an onomasiological theory of English word-formation". *Onomasiology Online* 2. 1–42. <http://www.ku-eichstaett.de/SLF/ EngluVglSW/stekauer1011.pdf> (Last accessed 1 April 2010.)
Štekauer, P. 2005. *Meaning predictability in word-formation: Novel, context-free naming units*. Amsterdam: Benjamins.
Stepanova, M.D. 1973. *Methoden der synchronen Wortschatzanalyse*. München: Max Hüber.
van Marle, J. 1985. *On the paradigmatic dimension of morphological creativity*. Dordrecht: Foris.
Verspoor, C. M. 1998. "Predictivity vs. stipulativity in the lexicon". In: *Proceedings of the Pacific Asia Conference on Language, Information, and Computation*, Singapore. 152–162.

Address correspondence to:
Jesús Fernández-Domínguez
University of Jaén
Campus Las Lagunillas, s.n.
23071 Jaén
Spain
jesusferdom@gmail.com