Some Observations on the Structure, Type Frequencies and Spelling of English Compounds

Stanimir Rakić

This paper deals with the structure, type frequencies and spelling of English noun – noun compounds.[1] *On the basis of a corpus extracted from LDCE (2000), I note some regularities governing the structure, type frequencies and spelling of this type of English compounds. My main point is that the type frequencies and spelling of English compounds depend on the complexity of their constituents. The relevant generalization seems to be that spellers tend to insert a space in noun – noun compounds if any of its constituents is morphologically complex. In respect to the first constituent this tendency is particularly strong, so that solid compouds in English are overwhelmingly written with a monomorphemic first constituent. The exceptions to this generalization are not numerous and almost all can be accounted for in a principled way. The compound types which are easier to process are more frequent, and are also more often spelled solid. According to the proposed analysis of English compounds, solid compounds seem to differ from open compounds in four important features: spelling, morphological structure, type frequency and productivity.*

Keywords: *English compounds, constituent structure, type frequency, spelling, parsability, productivity.*

1. Introduction

It has often been noted in linguistic literature that noun – noun compounding is a very productive word-formation process in English. Thus, Bauer and Huddleston (2002, 1647) note that noun - noun compounding "is by far the most productive kind of compounding in English, and indeed the most productive kind of word-formation." We come across similar statements in Plag (2003: 145), Lieber (1993),[2] Katamba (1994: 74) and Bauer (2001: 117), usually based on the observation that new compounds are easily formed and accepted in English.[3] However, different opinions are encountered in psycholinguistic literature, due to the fact that there are few rules governing the compounding of two lexemes in English. "In English, lexicographers find new compounds by examining popular usage – that is, words used together relatively often to denote a specific concept" (Inhoff et al. 2008). Juhasz et al. (2003: 224) also remarked that "there are no systematic rules in English for the compounding of two free lexemes", which presumably exist in other languages. Inhoff et al. (2000: 24) point out that most Germanic languages other than English "permit the generative compounding of words" into novel compounds that are "written without interword spaces." The lack of systematic rules for compounding in English is cited as a possible reason for the variability of the spelling of compounds: lexicographers simply pick the spelling most frequently found in written sources. Thus conventional spelling is established for many compounds, while the spelling of others continues to vary.

Most authors who have studied compounding in English have mainly dealt with compounds whose constituents are monomorphemic words.[4] In this paper I argue that there are important constraints governing the possible structure of compounds whose constituents are complex and which have so far have passed unnoticed.[5] My observations are based on material from the LDCE (2000). However, I do not believe that this fact impairs my results,

as the spoken language is usually simpler than the written one. It is therefore not plausible that there are compound structures in the spoken language which are not represented in the corpus extracted from the LDCE (2000).[6] The extracted corpus shows that the number of possible compound types drastically decreases if the constituents are complex.[7] It is also possible to see that with increased complexity of the constituents, fewer compounds are written solid, and they increasingly must be written open or hyphenated.[8] It is true that greater complexity of constituents goes hand in hand with increased length, but I will argue that the complexity of constituents is also active as an independent factor. I therefore argue against the thesis that English spelling is "extremely inconsistent in dealing with noun + noun collocations" (Bauer 1998: 69).

The corpus excerpted from the LDCE (2000) contains 5270 compounds. In the parsing of compounds we had in view paradigmatic relations between nouns which have the same basis (cf. Booij 1996). So we parse the noun *ambulance* into *ambul+ance* because the words *ambulant* adj. and *ambulate* v. derive respectively from the suffixes *-ant* and *-ate.* Similarly, the noun *militia* can be parsed into *milit+ia* because the words *militant, militate* and *military*, derive from the suffixes *-ant, -ate* and *-ary* and have related meanings. The names of sciences like *optics* and *physics* can probably be treated as complex too because of their relationship to the adjectives *optical* and *physical*. Similarly, for *economics* and *politics* there are bound bases respectively in the nouns *economist* (or *economy*) and *politician* (or *policy*). If a noun is derived both with a prefix and suffix, I assume that its structure is binary, e.g. the noun *permission* is decomposed into *permit+ion,* where the allomorphs must also be taken into account. The noun *enforcement* is decomposed into *enforce+ment,* and is understood as a suffixal derivation although the prefix *en-* is also involved in its derivation. The initial combining forms which appear as segments in nouns like *television* and *psychoanalysis* have been treated as prefixes in the analysis of compounds. Prćić (2005) has shown that there are important semantic, phonological, morphosyntactic and etymological differences between prefixes and initial combining forms in English. In this paper, however, most important is the fact that these forms similarly contribute to the morphological complexity of derived words. For nouns lexicalised with the plural *-s* (e.g. *arms, scissors, billiards)*[9] I have followed the traditional approach and tagged them as morphologically simple; I have treated in the same way words such as *cavalry* and *velocity* in which the final segments can be recognised as particular suffixes, but the other parts of these words do not exist independently. The nouns *huckleberry, Wednesday* and *iceberg,* in which one constituent is clearly recognised as an independent word, while the rest of the word is not clearly motivated, I have nonetheless classified as compounds – I thereby follow Fabb (1998), who classified such words as non-prototypical compounds. His argument was that the rest of these examples have lexical, not grammatical meaning.[10] Here I assume that the constituents of compounds can be simplex nouns, suffixal or prefixal derivatives, compounds or phrases.

The structure of this contribution is organised as follows. In section 2, I will review some recent studies on the spelling of compounds. In section 3, the statistics based on the extracted corpus shows how type frequency and the spelling of compounds depend on the complexity of constituents. In section 4, I note some regularities and constraints in the structure of compounds, their type frequency and spelling which can be observed in the distribution of compounds. In section 5, I analyse two crucial cases in more detail and show how some apparent exceptions can be accounted for. In section 6, I analyse the case of compounds containing nouns with the lexicalised plural *-s* (the so called 'pluralia tantum') as one of their constituents. In section 7, an alternative account of the noted exceptions is

tentatively provided capitalising on the notion of lexical frequency, and finally in section 8, I sum up the main results and indicate possible directions for further research of this topic.

2. A brief review of some recent analyses of the spelling of English compounds

It is well known that English compounds may be written in three different ways even in the case of having the same accent (Jespersen 1942: 136). Marchand (1969: 21) also noted "the complete lack of uniformity" in the spelling of English compounds. Quirk et al. (1985) made the same observation noting that some compounds may be written in all three ways: 'solid', 'hyphenated' or 'open'. Nonetheless they concluded that there is a progression from "open" to "solid" spelling "as a given compound becomes established." Inhoff et al. (2000) similarly observes that spaces are commonly deleted in compounds which are used relatively often, so that the place of residence on a farm is written as *farmhouse,* but an accident on a farm may be written as *farm accident* rather than *farmaccident.* They further note that less familiar compounds or novel compounds are rarely written without a space. Bauer (1998: 69), however, finds that there is no steady diachronic progress towards solid spelling based on "compound frequency and the age of the compound" since some familiar compounds are spelled with a space (e.g. *college degree),* and some new ones (e.g. *airside* from *The Oxford Dictionary of New Words*) are spelled as one orthographic word. As the spelling of compounds "depends on the taste and fancy of the speller it cannot represent a consistent linguistic judgment about the nature of the constructions". We find similar remarks in Juhasz et al. (2003: 224), who noticed that the spelling of compounds seems to be largely determined by popular usage, since lexicographers follow the spelling found in written sources. Thus, the spelling of some compounds becomes relatively standardised by convention, while the spelling of others may vacillate in all three ways (e.g. *life style, life-style* and *lifestyle).* De Jong et al. (2002) found that 19.21% of compounds in their corpus occurred in an alternative spelling. The most frequent variation in their corpus based on Celex was between open and hyphenated compounds.

Bauer notes, however, that the spelling of compounds is not absolutely without consistency: long words tend to be written separately, while short words are likely to be written together. The distinction in the number of letters is, however, not a structural linguistic distinction, but "depends on factors such as ease of perception for the reader." Bauer does not believe that such an observation can provide a valuable linguistic generalisation. I nevertheless believe that this remark of his is a hint pointing in the right direction. Without any doubt, the spelling of compounds is not governed only by linguistic structures: some pragmatic factors regulating communication between the speller and the reader must necessarily be present. Following Grice's communicative postulates the speller must try to make his message the most easily available to the reader (Levinson 1983: 102). It is easier to parse a shorter string of letters than a long one, and spaces in longer compounds make the task of the reader much easier. Inhoff et al. (2000), who have studied German trinomial compounds, have shown that spaces are more efficient signs of morpheme boundaries than any of the other possible linguistic or orthographic signs: consonant clusters, affixes or capital letters. The subjects respond faster to compounds whose members are separated by spaces than in any other way, although German compounds are conventionally spelled without spaces. That the compounds spelled with spaces are responded to faster than the concatenated ones has been confirmed in a number of recent experiments (cf. de Jong et

al. 2002, Juhasz et al. 2003, 2005, Huijer and Krott 2004). De Jong et al. (2002) have presented some evidence that the constituents of English open compounds are processed in a manner more similar to the processing of simplex words than the constituents of solid compounds, whereas Juhasz et al. (2005) have noticed that "inserting a space into a normally nonspaced compound" can significantly disrupt processing. This seems to happen in the case when the compound is conventionally spelled without spaces and its meaning is not fully compositional.[11]

It is true that the speller tends to insert spaces in longer compounds and to write shorter ones in a unified way, but nonetheless in the spelling of compounds there is more regularity than is usually assumed. In our corpus 82.17% of compounds with more than two syllables are written open or hyphenated, and only 17.83% without spaces, while only 37% of compounds with two syllables are written open or hyphenated, and 62.27% without spaces. Which disyllabic compounds are written with a space or hyphen, and which solid is largely determined by semantic factors, conventions and even phonotactic transitions.[12] The spelling of polysyllabic compounds is probably influenced by the same factors, but also by the morphemic structure of the constituents which in disyllabic compounds plays a minor role, since in these compounds only the addition of the suffix *-s* is possible. However, the compounds with more than three syllables are overwhelmingly written open so that the impact of the morphological structure of constituents becomes less evident with the greater number of syllables. In section 5, the analysis shows that the morphological structure of constituents has an important impact on the spelling of compounds in the cases for which we can assume that the number of syllables is not considerably different.

In this paper I show in particular that compounds in which the first constituent is a suffixal derivative tend more often to be written open than those in which the second constituent is a suffixal derivative. The exceptions to this generalisation are not numerous, and I will argue that they can be explained in a principled way. Furthermore, it is shown that compounds whose constituents are prefixal derivatives are almost always written open, and this happens regardless of whether the first or the second constituent is built in that way. The fact is, therefore, that compounds with constituents containing prefixes are spelled differently from those whose constituents contain suffixes, and the reason for this difference probably lies in the different way of processing words with prefixes and those with suffixes (cf. Taft and Forster 1975, Cutler et al. 1985, Hay 2002: 94-97). The morphemic structure of constituents of polysyllabic compounds is an independent factor influencing both the spelling and type frequency of compounds. The complexity of compound constituents seems to impose some kind of limit to the frequency of certain compound types.

3. The statistics of compound types and spelling

In the corpus of 5270 noun + noun compounds extracted from the LDCE (2000), I was able to note the following distribution of compounds in view of their constituent structure:

(1a) Stem + Stem ('Stem' denotes a simplex noun throughout this paper).
There are 3457 examples of such compounds: 1798 of them are written open, 1576 solid, and 83 hyphenated. Here belong also 39 compounds of the type *arms race* or *bar billiards,* the overwhelming majority of which are written open (for discussion see section 6 below).

(b) Stem + (Stem + Suffix)

There are 718 examples of such compounds (*adult education, airwaves, aid worker, air conditioning, air freshener, air-hostess, bartender, baby-minder, beachcomber*…). 426 are written open, 206 solid and 86 hyphenated.

(c) (Stem + Suffix) + Stem

There are 655 examples of such compounds (e.g. *Nativity play, Yorkshire pudding, absentee ballot, aircraftwoman, amusement arcade, apartment block, blotting-paper, salesclerk* …). 575 examples are written open, 73 solid and 7 hyphenated. The compounds *blotting-paper, feeding-bottle, meeting-house, piggy-bank, question-master, stepping-stone, brigadier-general* are written with a hyphen.

(d) (Stem + Suffix) + (Stem + Suffix)

There are 116 examples of such compounds (e.g. *Virginia creeper, absorption costing, action stations, advertising agency, aircraft carrier, ambulance service, apartment building, assistant professor, sales representatives, vacuum cleaner,*[13] etc.), of which 107 are written open, 7 with a hyphen (*fender-bender, decision-making, jiggery-pokery, owner-occupier*, *stretcher-bearer, wheeler-dealer, washer-drier)*, and 2 solid (*thanksgiving, whippersnapper*).

(e) ((Prf + Stem) + Stem) or ((Prf + Stem) + (Stem+Suffix))

There are 26 examples of such compounds (e.g. *cauliflower cheese, deposit account, dispatch box, exchange rate, import duty, midlife crisis, midnight sun, reserve price, peroxide blonde, relay race, submachine gun, telegraph line, telephone book,* etc.). Only 2 of them are not written open (*mischief-maker* and *midshipman*). In 3 examples the second constituent is a suffixal derivative (*dispatch rider, mischief-maker, telephone directory*), and in the rest of the 23 examples a simplex noun.

(f) ((Stem + (Prf + Stem)) or ((Stem+ Suffix) + (Prf + Stem))

There are altogether 29 examples of such compounds (e.g. *air vice-marshal, beauty contest, cash discount, day return, drug misuse, glove compartment,*[14] *information superhighway, market research, user interface,* etc.). Of them only 2 compounds are written with a hyphen (*part-exchange, radio-telephone).*

In 5 examples the first constituent is a suffixal derivative (*action replay, eating disorder, freezing compartment, information superhighway, and user interface*), and in the rest of 24 compounds a simplex noun.

(g) (Prf + Stem) + (Prf + Stem)

There is only 1 compound whose constituents are both prefixal derivatives. This is the compound *telephone exchange.*

(h) (Stem + Stem) + N, where N denotes a simplex noun or a derivative.

There are 152 examples of compounds of this type, and of that number 118 compounds have a simplex second constituent (e.g. *Scotland Yard, air traffic control, apple pie bed, baseball cap, backdown track, bulldog clip, big horn ship, breakdown*

track, buckwheat flour, butterfly nut, capital gains tax, cardboard city, catbird seat, cocktail lounge, cut-throat razor, dustbin man, etc.).

34 compounds have a derived second constituent (e.g. *backroom boys, Sunday driver, airtraffic controller, backseat driver, ballroom dancing, buckwheat cakes, caretaker government, cavity wall insulation, tenpin bowling*...). In this set all second components are suffixal derivatives.

From all these compounds only 6 are written solid (*highwayman, holidaymaker, longshoreman, underclassman, upperclasswoman, upperclassman*). It is conspicuous that all these compounds except one – *holidaymaker*, are those in which the words *man* or *woman* are a second component.

(i) N+ (Stem + Stem), where N denotes a simplex noun or a derivative.

There are altogether 58 examples of this type of compound. 45 of these compounds have a simplex first component (e.g. *Molotov cocktail, Venus flytrap, air chief marshal, bank holiday, bubble jet printer, banner headline, cable railway, fairy god-mother, flag football, gas permeable lens, germ warfare, grade point average, hormone replacement therapy, horse chestnut, roll-top desk,* etc.).

There are 13 compounds with a derived first component, all cited here (*absentee landlord, adventure playground, computer dating agency, consumer price index, daddy longlegs, drilling platform, government health warning, moving staircase, package holiday, safety –deposit box, singer-songwriter, sleeping policeman, Mothering Sunday).*

Altogether 7 compounds are written with a hyphen (*car-boot sale, dot-matrix printer, penny-halfpenny, radio-cassette player, roll-top desk, singer-songwriter, safety-deposit box*), and all the rest are written open. The first component, simplex or derived, cannot be written solid with the second component. In the previous group, the second component is written solid in 6 examples, but in 5 cases these are the nouns *man* or *woman,* and the remaining one is the compound *holidaymaker* in which the first component is to a great extent a lexicalised noun *holiday*. The difference between the two groups in the distribution of complex words and spelling is evident.

(j) (((N+N)+N)+N)…

In our corpus there is just 1 compound with recursive structure – *daylight saving time,* the structure of which is (((N+N)+N)+N). In the generative framework, such constructions are often cited as illustrations of the claim that compounding is in principle recursive, i.e. that compounds may surface as a part of other compounds even in an infinite sequence. Selkirk (1982) gives the following examples:

I) bathroom
II) ((bathroom) + (towel rack))
III) (((bathroom) + (towel rack)) + designer)
IV) ((((bathroom) + (towel rack)) + designer) + training)

Such repeatedly inserted compounds are regularly written separately. Because of the difficulties in processing complex forms like (III) and (IV), compounds of that type are rarely used in spoken language – they are not convenient naming units (cf. Štekauer 2001), and usually they are not found in dictionaries.

(k) NP + X,[15] where X is a simplex noun, derivative or compound.

Phrases may also appear as constituents of English compounds, although it is usually added that only lexicalised ones can take such a role (Plag 2003). Altogether, there are 55 compounds of this type in our corpus. The other component is a simplex noun in 48 examples (e.g. *bird's-eye view, devil's-food cake, dry-stone wall, fair-weather friend, five-star general, four-letter word,*[16] *heart-lung machine, hot-water bottle, kitchen sink drama, left-hand drive, lonely hearts club, mad cow disease, minimum security prison,* etc.).

In 7 examples the second component is a derivative (*quick-change artist, hand-eye coordination, open door policy, scorched earth policy, tenpin bowling, supply side economics, time-space continuum*).[17]

All compounds of this type are written open, except for the compound *rag-and-bone-man*, in which the second component *man* is connected with a hyphen.

(l) X + NP, where X may be a simplex noun, derivation or compound.

There are only two compounds of this type. With a simplex first component there is one compound (*Jack the Lad)* and one with a derivative (*beggar-my-neighbour*). In the last example the components are written with a hyphen. In our corpus there are no compounds in which both constituents are phrases.

4. The observations

The analysis of the distribution of English compounds in the corpus extracted from the LDCE enables me to note some important constraints governing the subsets of English compounds and also to note some definite tendencies in the spelling of English compounds with complex constituents. It is not difficult to notice that in the subsets (1a) – (1l) the number of compound types progressively decreases as the complexity of their components increases. There are 3457 compounds whose components are both simplex nouns; the number of compounds in which one component is a suffixal derivative is much smaller –1373, and even lower is the number of compounds in which both components are suffixal derivatives, only 116. A prefixal derivative in one component appears in 55 compounds, but components are both prefixal derivatives in only one compound. There are altogether 210 compounds in which one component is also a compound, but there is only one compound in which both components are compounds. 57 compounds include a phrase in one of their components, but none have phrases in both components.

For compounds which include prefixal derivatives in one component, the number in which the other compound is a simplex noun is greater than the number in which that component is a suffixal derivation. Similar relations are found in compounds having a compound or phrase as one component. In those in which one component is also a compound, the other component in the majority of cases is a simplex noun, in a smaller number of cases a suffixal derivative, and in our corpus, never a prefixal derivative. In the compounds in which one component is a phrase, the other component in a majority of the cases is a simplex noun, in a smaller number of cases a suffixal derivative, in one example a compound, but never a prefixal derivative. It seems that the inclusion of prefixal derivatives makes a compound structure more complicated than the inclusion of suffixal derivatives. It is not difficult to

understand that ease of processing is the reason why the number of compound types drastically decreases if their constituents become more complex. The compound types with a complex structure are naturally more difficult to process, and speakers as well as spellers obviously tend to avoid them. The compound types which are easier to process are therefore also more frequent. In a similar way, Cutler et al. (1985) have argued that the preference of language users to process stems before affixes has brought about as a consequence the wider use of suffixes than prefixes in world languages (Hay 2003: 10).[18]

The complexity of compound constituents has important consequences not only for the frequency of compound types, but also for the spelling of compounds. The greatest number of compounds are spelled solid when both components are simplex nouns (45.59%), and this percentage drastically decreases if components have a more complex structure. In our corpus, the compounds in which one component is a prefixal derivative or a compound are very rarely spelled solid, and if one component is a phrase, no compound is spelled solid. The statistics of the previous section shows that the decrease in solid spelling parallels the lowering of the number of compound types. Ease of morphological processing obviously influences both the number of compound types and their spelling. That response to compounds spelled with spaces is faster than to those without spaces has been confirmed in a number of recent experiments (cf. de Jong et al. 2002, Juhasz et al. 2003, 2005, Huijer and Krott 2004). For ease of perception, people writing in English tend to spell compounds with a space or hyphen if they are long, and this most often means when they contain complex constituents. The variability in spelling follows mainly for pragmatic reasons – spellers must weigh the pros and cons for open or solid spelling in circumstances when conventional spelling is not fixed, and they do not always agree. The statistics in the previous section shows that spellers tend to insert a space in noun – noun compounds if any of its constituents is morphologically complex. In respect to the first constituent this tendency is particularly strong, so that solid components in English are overwhelmingly written with a monomorphemic first constituent. It is true, however, that the spelling of compounds also depends on the number of syllables. In the next section I analyse in more detail the influence of an unequal distribution of suffixal derivatives in compound constituents on the spelling of compounds, and thereby show the independent effects of morphological complexity.

5. The two crucial cases

Our survey shows that the spelling of compounds depends on the complexity of the constituents. Since compounds with complex constituents are usually longer, it is important to discuss cases in which the complexity of constituents makes the crucial difference. Bauer (1998) has already noted that longer words tend to be written separately, regardless of stress pattern, while short words are more likely to be written together. In this section we point to the cases in which the crucial difference is exactly the complexity of constituents. The general rule is that the more complex the constituents, the more frequently compounds are written open or with a hyphen, but the way of spelling also depends on the kind of complexity. We can single out three factors which seem to favour the open spelling of compounds: prefixes contained in constituents, the compound structure of constituents and the inner position of suffixes. We can show the impact of the inner position of suffixes by comparing subsets (1b) and (1c) presented above. There are 718 compounds with the structure Stem + (Stem+Suffix) of which 426 are spelled open, 206 solid and 86 hyphenated. The percentage of the

compounds written open is 59.33%, those written solid 28.69%, and hyphenated 11.98%. The corresponding numbers of the compounds with the structure (Stem + Suffix) + Stem are respectively 655, 575, 73 and 7 which yield the percentages 87.79%, 11.14% and 1.07%. These numbers and percentages are shown in Table 1:

Compound	Sum	open	Solid	hyphenated
N+(N+Suf)	718 = 100%	426 = 59.33%	206 = 28.69%	86 = 11.98%
(N+Suf)+N	655 = 100%	575 = 87.79%	73 = 11.14%	7 = 1.07%

Table 1

The great difference between these percentages shows that the structure of constituents decisively influences the way in which compounds are written. We assume that the compounds in subsets (1b) and (1c) do not differ considerably in length. This means that the complexity of constituents is an independent factor in the spelling of compounds. It is crucial that the difference in the position of prefixes does not make the same difference in the spelling of compounds (cf. (1e) and (1f) above). Note also the huge difference in the percentages of the hyphenated compounds in subsets (1b) and (1c). The number of hyphenated compounds lowers in the same direction as the number of solid compounds as we pass from set (1b) to set (1c).

Further analysis shows that the number of compounds written solid in subset c) is not at all secure because components of some compounds may undergo different interpretations. The entire list of these compounds is given in (2).

(2) aircraftman, aircraftwoman, ambulanceman, assemblyman, assemblywoman, backwoodsman, batsman, bridesmaid, businessman, businesswoman, carriageway, chatterbox, chatterline, checkerboard, clansman, clanswoman, clapperboard, cockscomb, companionway, craftsman, craftswoman, cranesbill, draftsman, draughtsman, fieldsman, fisherman, frontiersman, groomsman, groundsman, guardsman, halterneck, helmsman, herdsman, huntsman, jobsworth, kinsfolk, kinsman, kinswoman, laboratoryman, lambswool, linesman, marksman, militiaman, nurseryman, oarsman, oarswoman, passageway, pointsman, pokerwork, rubberneck, salesclerk, salesgirl, salesman, salesperson, saleswoman, serviceman, servicewoman, sportsman, sportswear, sportswoman, statesman, steersman, swordsman, townspeople, townsfolk, tradesman, trawlerman, tribesman, tribeswoman, washerwoman, woodsman, yachtsman, yachtswoman.

An interesting feature of the compounds in (2) is that some forms appear repeatedly as a second component. In literature, the items *-man, -woman, -person, -work* and *-ware* are sometimes labelled as ‘combining forms’ (the COED). To this set we can add the string *-way* which in the COED has been tagged surprisingly as a full suffix. If we accept such a classification, our thesis becomes much stronger. As these forms surface 57 times as second components in (2), the actual number of the compounds written solid shrinks to 16, since combining forms take part in derivation, not in compounding. This would reduce the percentage of compounds written solid to 2.68% in subset (1c),[19] a result that convincingly shows the role of constituent complexity in the spelling of compounds. In addition, a similar

relationship surfaces in subset (1g), where, apart from the word *holidaymaker,* only the compounds in which the second constituents are the items *man* and *woman* are written solid. Whether we accept the new count or not, it is fairly clear that the structure of constituents heavily influences the way in which English compounds are written. The use of the term 'combining forms' is justified in the context of our analysis because it points to the particular character of the compounds in (2), in which some lexical units repeatedly appear as a second constituent.

In (2) the status of *-s* is also questionable. I have treated it as a full-fledged suffix in order to make the opposite thesis stronger. The inflectional *-s* can be understood as a sign of plurality or possession, and in many cases it is prone to lexicalize with its host. MacWhinney and Snow (1985) have shown that "the plural /s/ morpheme is a perfect predictor of word finality". That means that the plural -s makes decomposition easier. The opposite pattern holds for the possessive s – it is always followed by a second noun. It follows that compounds with plural *-s* must be written solid, if they are not intended to be understood as genitive phrases. One can, therefore, tentatively claim that the solid spelling of compounds with the plural *-s* follows from independent principles which would imply a further reduction of the number of exceptions in (2). In the compounds *bridesmaid* and *groomsman*, *-s* is ambiguous between plural and possessive meaning, and solid spelling promotes the plural meaning into the first plan – these compounds primarily have general, collective meaning, while possessive meaning will imply some kind of reference. Still, some compounds with plural *-s* are spelled with a blank:

(3) drinks machine, honours degree, sports centre, systems analyst, trades union, trials bike.

It is, however, easy to see that in the examples (3) no genitive interpretation is possible as the first components are not animate; these examples simply follow the more general rule: the second constituents in (3) are not combining forms.

Let us now look finally at the other exceptions in (2), those in which first constituents do not end in *-s* nor have 'combining forms' as second constituents. These are the following compounds:

(4) chatterbox, chatterline, checkerboard, clapperboard, halterneck, rubberneck.

In all of these examples the first constituents contain the suffix *-er,* the pronunciation of which is /ə/. According to the well-founded distinctions in Hay (2003), this is a 'legal' phonotactic transition. This means that the first constituents in (4) are, at least at prelexical level, not perceived as parsable nouns,[20] a fact which partially accounts for the generations of the compounds in (4). Some of the first constituents in (4) could to a native speaker have looked similar to simplex nouns. The first constituents in (4) are of course complex words, but ones which are close to the periphery of this category.[21] Several examples in (2) derived from combining forms also have similar first constituents (e.g. *ambulanceman, assemblyman, assemblywoman, companionway, laboratoryman, militiaman, nurseryman, trawlerman,* etc.). This is a case of rule conspiracy which could be the subject of further detailed examination, into which we could not venture here.

Our conclusion however can be further generalised since other compounds with a complex first constituent are also rarely spelled solid – the only examples being the compounds noted in (5):

(5) thanksgiving, whippersnapper, holidaymaker

The compounds in (5) differ from the compounds in (2) in that the second constituent is a suffixal derivative. Now we can calculate the percentage of the solid-spelled compounds taking into account all compounds with a complex first constituent. Again, we do not count the examples in (1h) and (1e) with the formants *man* and *woman* as second constituents. There are altogether 1028 compounds in which the first constituent is a complex word, and between them 19 compounds are written solid. Simple calculus shows that 1.84% of compounds with a complex first constituent are written solid. This means that there is a strong tendency to write almost all compounds open in which the first constituent is a complex noun.

We should admit that the use of the notion of 'combining forms' in this section is somewhat controversial. Many authors reserve the term 'combining forms' for formants borrowed from Latin and Greek (cf. Prćić 2005), and reserve the alternative term 'semi-suffixes' for the formants which massively appear as second constituents in compounds, show to a certain extent reduction of meaning, and, occasionally, pronunciation. There is also no general agreement as to which formants should be assigned to this class. Marchand (1969) applies this term to the formants *-like, -worthy, -monger, -way* and *-ways, -wright.* Allen (1978) treats as this kind of suffix the formants *-man* /mən/, *-land* /lənd/ and *-berry /bə*r/ in words like *fireman, highland* and *raspberry* in which the vowels /æ/ and /e/ have been reduced to schwa. The LDCE does not include in this class the formant *way*, although in the LDCE this formant occurs much more often as a second part of word combinations than as a first. The combining forms (or semi-affixes) have been recently analysed by Booij (2005) from the position of the theory of grammaticalization and construction grammar. The rise of semi-affixes is a typical process of grammaticalization in that content words are becoming grammatical morphemes. When they are used as suffixes their meaning is reduced, but they can still be used independently with "a greater range of meaning". According to this interpretation, words derived with combining forms are special constructions occupying an intermediate position between compounds and derivatives. In section 7, I tentatively propose an account which does not involve the notion of combining forms.

6. The problem of nouns with lexicalised plural

It is interesting to examine how nouns with lexicalised plural in *-s* (the so called 'pluralia tantum') behave in compounds when they figure there as compound constituents. According to our taxonomy which follows the traditional one, such nouns are classified as simplex nouns, and accordingly, should have no restrictions on occurring both in open and solid compounds. In our corpus there are 39 compounds which include in one of their constituents a noun with lexicalised plural in *-s*. In 30 compounds *-s* occurs in the second constituent, and in the other 9 in the first constituent. 11 compounds are written solid, but only in one of them does *-s* take the inner position. This distribution is shown in (6):

(6) (a) clothesline

(b) bedclothes
eyeglasses
gasworks
handcuffs
headquarters
nightclothes
soapsuds .
steelworks .
sunglasses
waterworks

Distribution (6) demonstrates that the compound structure much more easily tolerates nouns with lexicalised *-s* in the second than in the first member, which suggests that nouns with lexicalised *-s* are treated as complex nouns. Except for *clothesline,* all compounds with *-s* in the first constituent are written open in our corpus:

(7) arms control, arms race, clothes basket, clothes hanger, clothes brush, clothes horse, clothes peg, gallows humour.[22]

But surprisingly, in some cases *-s* in the middle position can fall away. If a noun with *-s* happens to occur in the first position, it may lose the suffix! We see *-s* truncated not only in (8a) in solid spelling, but also in (8b) in open spelling:

(8a) oatmeal n. < oats n.pl.
knickerbockers n. < knickers n.pl.
scissorbill n. < scissors n.pl.

(b) trouser suit, trouser press < trousers n.pl.
pyjama bottoms < pyjamas n.pl.
billiard table < billiards n.pl.
clipper chip < clippers n.pl.
Balkan states < the Balkans n.pl.
pincer movement < pincers n.pl

(c) pantyhose n
pantyliner n

In (8c) we find a replacement of -s with /*I*/ (in spelling *y*), which obviously serves as a linking element. Interestingly, in BE the word *pants* may be used as a predicative adjective (e.g. *The concert was pants*), but not before nouns as a modifier! In the LDCE the words *trouser, pyjama, clipper, pincer* are interpreted as attributive adjectives, i.e. adjectives which occur only before nouns. Such tagging, however, is unmotivated since nouns in English quite freely occur as modifiers in compounds. Now the truncation of the plural *-s* in 8) does not seem unmotivated after we have seen that the distribution of suffixal derivatives is quite limited in the modifier position in compounds. The alternation *-s/ø* in (8a) and (8b), or *-s* replacement with *y* in (8c) are now much easier to understand.[23]

This evidence now raises the question of whether or not I was justified in the first place in tagging the nouns with lexicalised plural *-s* as morphologically simple. In doing so, I have followed the traditional classification which was obviously motivated by the semantic opacity of such derivatives.[24] Morphologically, however, the plural *-s* is often easily parsable because in most cases it builds improbable phonotactic transitions (Hay 2003). This would imply that the nouns with lexicalised plural *-s* are morphologically complex in cases when the suffix *-s* builds an improbable phonotactic transition. In that case, the suffix -s falls away more easily from the attributive position, insomuch as the plural meaning of modifiers in

compounds is usually not required in English. The result of the analysis is quite obvious, although not an expected one: nouns with lexicalized plural *-s* can be morphologically complex in different degrees. Here the assumed tagging of nouns with lexicalised *-s* has not significantly harmed our survey because the examples in (8) are invisible for our statistics, since the modifiers in (8b) have been classified in the LDCE as adjectives. As we mentioned before (cf. footnote 10), a consistently applied analysis can result in appropriate corrections.

Examples (6), (7) and (8) clearly show that there is quite a strict restriction on the occurrence of plural *-s* in the inner position of solid compounds. This is obviously a special case of the restriction holding in general for all suffixes, and therefore independent evidence for that restriction.

The well-known problem of plural nouns in the first position in English compounds may have some connection with the fact that plural *-s,* in general, is easily parsable. Are the compounds in (2), (3) and (7) possible counterexamples to the claim that easy parsability of plural *-s* contributes to its loss in compounds? Note however that in (2) *-s* appears mainly before combining forms, or before nouns which could be assessed as being half-way towards acquiring such a status. Already Allen (1978) wondered why *-s* surfaces only before the formants *man* and *woman*, which she tagged as second level suffixes. Also in (7) *-s* has a very restricted distribution – it appears after the sonants /m/, /w/ and the fricative /D/. Whether *-s* falls away or not may depend on the kind of transitions it builds, but also of its meaning. Recall that Sproat (1985) has argued that *–s* can keep its position inside compounds if the plural nouns have collective meaning, and this is a condition which seems to be fulfilled in examples (2), (3) and (7). Yet the phonotactic conditions in (3) seem to be different, showing that the well-known problem of the plural in English compounds requires much more elaboration than the few passing remarks that I can make here.

7. An alternative account – the effects of lexical frequency

In this section, I will tentatively try to provide an alternative account of examples 2) on the basis of evidence which is now available in psycholinguistic literature. This account is mainly based on the notion of lexical frequency. A large body of evidence has accumulated in psycholinguistic research in recent years showing the impact of lexical frequency on speech perception, parsability and production. Hay (2003: 5) notes that "less acoustic information is required to identify high-frequency words than low-frequency words, and lexical decision times are negatively correlated with lexical frequency."[25] Lexical frequency plays a particularly important role in the widely accepted dual-route model of lexical access to morphologically complex words. In this model a complex word may be accessed directly as a whole or decomposed via its parts. One of the main factors influencing which route wins is lexical frequency – if the derived word is more frequent than its constituents, the whole-word route has a good chance to win, but otherwise the decomposition route has the advantage. Applied in a simplified way to the processing of compounds, the race-model of lexical access would almost always favour the decomposition route because compound words in English are rarely more frequent than their parts. For example, *cocktail* – a very common word, has an adjusted frequency of 14.38, while the adjusted frequency of *tail* is 19.55 in Francis and Kučera's (1982) ranking. *Cranesbill* from (2) also drops out of the same list, which means that its adjusted frequency is less than 5, and for *bill* this frequency is estimated to be 100.13. However, another factor also seems to matter – the transparency of compounds. On the basis

of the assembled evidence, Frisson et al. (2008) suggest that words are automatically morphologically decomposed early at the visual processing stage, but at a later stage, when constituents have to be stitched together again, semantic transparency reaffirms its influence.[26] A number of studies have shown that the frequency of constituents plays an important role in the processing of compounds in English (Andrews, Miller and Rayner 2004, Juhasz, Inhoff and Rayner 2005, Juhasz, Starr, Inhoff and Placke 2003), indicating that the decomposition of compounds in constituents actually takes place, if any of the constituents is a high-frequency word. In particular, experiments have shown that lexical decision time for presented compounds and fixation times during reading are shorter when a compound's constituents are high-frequency lexemes. Some studies, however, have also found that the frequency of the whole compound plays a role, presumably at a later stage of processing (Bertram and Hyönä 2003). According to Taft (2004) this is especially the case when the meaning of the compound is not fully transparent, but deviates somehow from the composed meaning of its elements. The study of Juhasz et al. (2005) provides further support for this view.

For our problem, the results of the studies of Juhasz et al. (2003) and Inhoff et al. (2008) are particularly important. Juhasz and her colleagues examined the processing of solid English compounds in naming, lexical decision and sentence-reading tasks manipulating beginning and ending lexeme frequencies. In all the experiments they found that the processing of compounds was more effective when the ending lexeme was a high-frequency word. The effects of the beginning lexeme emerge in lexical decision tasks only when the ending lexeme is a low-frequency word. The authors conclude that the role of the beginning lexeme "is limited compared with that of the ending lexeme, assumedly because compound and ending lexeme meanings tend to converge" (ibid.: 244).

Their finding that the use of the ending constituent is the most influential part of compound recognition conflicts with van Jaarsveld and Rattink's (1988) study of Dutch compounds which revealed a larger beginning constituent effect. Juhasz et al. (2003) resolve this apparent inconsistency by noting that Jaarsveld and Rattink were mainly dealing with novel compounds. In one experiment in which they contrasted novel and lexicalised compounds, they also found a larger ending constituent effect. The conclusion is, therefore, that the ending constituent effect surfaces only in the recognition of lexicalised compounds, not of novel compounds. The results of Juhasz et al.'s (2003) work were further supported by Inhoff et al.'s (2008) study which found a larger effect for dominant constituents in the experiments with solid dilexemic English compounds. In all three tasks – lexical decision, naming, and sentence reading – the effect of the dominant constituent, regardless of whether the beginning or the ending, was confirmed. Similar findings have also been noted in the study of Libben et al. (2003).

The conditions assumed in the studies of Juhasz et al.(2003) and Inhoff et al. (2008) are very well applicable to the examples in (2) as they are common, lexicalised compounds, and almost all have transparent meaning with a dominant and relatively high-frequent second constituent. In (9), the second constituents of the compounds in (2) are ordered according to the frequencies given in Francis & Kučera's (1982) rank list which contains 5996 words out of 35996 words contained in the corpus.

(9) man (1,954.28), way (1,005.38), people (867.35), work (662.45), woman (404.83), foot (323.93), girl (320.58), person (289.47), board (225.81), bill (100.14), box

(69.64), neck (51.08), folk (43.91), maid (28.63), clerk (25.66), worth (12.94), wool (5.40), comb (6), wear (3).

The last two items in (9), *comb* and *wear,* are not included in the rank list of Francis and Kučera, which only contains words whose adjusted frequency is greater or equal to 5.[27] The last noun from (2) included in Francis and Kučera's rank list is the noun *wool.* In fact, this noun is the 2396th noun on this list, which confirms that all the nouns in (9), except perhaps the last two, are relatively very frequent nouns.[28] In (9) therefore, only the metaphorical compounds *chatterbox* and *cranesbill* deviate from the conditions required by Juhasz et al. (2003) by not being transparent, in addition to the last two nouns in the rank list of Francis and Kučera – *comb* and *wear.* This alternative explanation is therefore much more general than the one given in section 5 and, importantly, does not involve the somewhat controversial notion of combining forms.[29] It shows that the speller need not evade the solid spelling of words in (2) since the transparency and high-frequency of the dominant, last constituents make their recognition easy enough. High-frequency words are easily recognised, and presumably this fact gives the decomposition route an advantage similar to the one occurring in open spelling. This explanation seems to be very plausible, and is solidly based on the findings of Juhasz et al. (2003), Libben et al. (2003), Inhoff et al. (2008) and other recent psycholinguistic studies. Some incongruence of results, still present, may hopefully be resolved in further research.

8. Concluding remarks

The statistics of the compound sets (1a) – (1l) demonstrate that the type frequencies of English compounds as well as their spelling depend to a great extent on the structure of compounds and on the complexity of their constituents. It is not difficult to notice that both the type frequency of English compounds and their spelling are negatively correlated with the complexity of compound constituents – the number of compound types progressively decreases and they are more frequently written open as the complexity of their components increases. Three factors seem to favour the open writing of compounds: the prefixes contained in constituents, the compound structure of the constituents and the inner position of suffixes. In particular, there is a strong tendency toward open spelling if the suffixes added to constituents take the inner position in compounds. Interestingly, no substantial difference seems to depend on the position of prefixes – the compound words with prefixes tend to be spelled open regardless of the position the prefixes take.

If the constituents are morphologically complex, the decoding of compounds is a more demanding task, and the speller naturally tends to alleviate this difficulty for the reader by choosing open spelling. This account seems to be compliant with the recent studies of Inhoff et al. (2000), de Jong (2002), Juhasz (2003, 2005), Huijer & Krott (2004) and Frisson et al. (2008) according to which open spelling of compounds provides easier access to constituent lexemes and thereby enhances the parsability and understanding of compounds. Variability in spelling follows mainly for pragmatic reasons – the spellers must weigh the pros and cons for open or solid spelling in circumstances when conventional spelling is not always fixed, and they do not always agree. The relevant generalisation seems to be that spellers tend to insert a space in noun – noun compounds if any of its constituents are morphologically complex. In respect to the first constituent this tendency is particularly strong, so that solid compounds in

English are overwhelmingly written with a monomorphemic first constituent. The exceptions to this generalisation are not numerous and cluster into a clearly defined pattern – the considerable majority of them have as a second constituent some of the items *-man, -woman, -person, -way, work* and *-ware,* which are sometimes labelled in literature as 'combining forms' (the COED). In section 5, I tried to show that the exceptions to this generalization can be accounted for in a principled manner. Another, apparently more general account has been provided in section 7.

In section 6, I extended the analysis to nouns with lexicalized plural *-s* (the so called 'pluralia tantum') which are traditionally understood as simplex nouns.[30] The analysis shows that nouns with *-s* mainly occur as a second constituent in compounds, and considerably less as the first constituent, and in our corpus in only one example as a first constituent in solid spelling. However, there are several compounds, both open and solid, in which pluralia tantum nouns occur without *-s*. In two such examples, *-s* is replaced with /*i*/, in spelling *y* (*pantyhose, pantyliner*). These facts suggest that the traditional classification of these nouns as simplex nouns is not entirely correct. Some of them are more parsable than others and therefore more complex (cf. Hay and Baayen 2005). However one might decide to treat the nouns with a lexicalized plural *-s,* their behaviour in compounds is independent evidence for the thesis that the occurrence of suffixes in the inner position of solid compounds is quite restricted. Whether our findings about the restrictions on the distribution of suffixes in compound constituents may have important implications for the well-known problem of plural in English compounds is left for further research. In section 7, an alternative account of exceptions noted in section 5 is tentatively offered. That account is based on the evidence available in psycholinguistic literature, and seems to offer a more general explanation than the previous one.

De Jong et al. (2002) have presented some evidence that the compounds spelled open are mainly accessed through their parts, and tentatively concluded that English open compounds have a different lexical representation than those written without a space, i.e. they have the representation which corresponds to 'orthographic phrases', and not to "orthographic words". If this is the case, the compounds with compositional meaning should favour open spelling since the decomposition route will be able to access the constituents directly, allowing quicker retrieval of the compound meaning. On the other hand, opaque or partly opaque compounds are accessed faster if they are spelled concatenated, since the decomposition route is less effective in that case (Taft 2004). This conclusion is quite consonant with the long-standing observation that lexicalised compounds tend to be written concatenated. Bauer's example *airside* shows that "no diachronic process of change in orthography" is necessary – the compound may be non-transparent from its beginning (Bauer 1998: 69). This just means that only some compounds become lexicalized as time goes by.

The analysis of the distribution of English compounds extracted from the LDCE enables us to note some important constraints governing the frequency of English compound types and to note some definite tendencies in their spelling. It is not difficult to notice that the number of compound types progressively decreases as the complexity of their components increases. There are 3456 compounds whose components are both simplex nouns; the number of compounds in which one component is a suffixal derivative is much smaller –1374, and even lower is the number of compounds in which both components are suffixal derivatives, only 116. A prefixal derivative in one component appears in 55 compounds, but both components are prefixal derivatives in only one compound. There are altogether 210 compounds in which one component is also a compound, but there is only one compound in

which both components are compounds. 57 compounds include a phrase in one of their components, but none have phrases in both components. Simultaneously with the increased complexity of the constituents, fewer compounds are written solid, and must increasingly be written open or hyphenated.

This analysis suggests that the human mental ability to process more complicated compound structures in the course of normal social intercourse is quite limited. Recent psycholinguistic research has confirmed that compounds written with a space are easier to parse and understand. As a visual cue – a space – cannot be supplied in speech, compound types with complex constituent structures, although quite frequent in some types of written English, are not used so often in spoken English (Miller 2006). For example, in our corpus there is just one compound with a recursive structure – *daylight saving time.* Without exception, compounds of this type are always spelled open – writing so the speller shows his/her willingness to make the reader's task easier.[31] This compound*,* the only recursive compound in the LDCE, cannot be found in the spoken component of the BNC. We should note, however, that compounds do not seem to be used frequently in English according to the frequency lists in Francis and Kučera (1982) and Leech et al. (2001). The reason for such ranking of compounds is that frequency lists do not count open compounds which are very common in written English, especially in the press, and solid compounds do not seem to be particularly productive in English (Inhoff et al. 2000, Juhasz 2003). Hence, we can conclude that solid compounds seem to differ from open compounds in four important features: spelling, morphological structure, type frequency and productivity. In spoken English, on the other hand, the counterparts of open compounds tend to realize only those compound types which have simpler constituent structure, although one must presume some variability depending on the genre. The behaviour of compounds in spoken English is still an area in which much more research is needed.

I argue, therefore, that the ease of morphological processing influences the distribution of complex words in English compounds and the way in which they are written. Compound types which are easier to process are more frequent, and are also more often spelled solid. In recent linguistics, there have already appeared analyses which proposed accounting for morphological facts by the way language is usually processed. For example, Cutler et al. (1985) argued that the preference of language users to process stems before affixes has as a result the wider use of suffixes than prefixes in world languages. Similarly, Hay argued that the ordering of affixes can be explained by the ease of parsability of particular affixes (cf. also Aarts 2004). The case of spelling is, however, more complicated – as a part of communicative acts, the account of it must involve pragmatic factors. The spelling can be explained only if we consider the options which the speller has in the given communicative situation. Following Grice's communicative postulates, the speller has to try to make the message most easily available to the reader, and normally adopts the way of spelling suitable for the message he/she wants to communicate. The variability of the spelling that we encounter with some compounds is connected with the fact that spellers do not necessarily make the same decisions. Our prediction, the verification of which is left for further research, is that compounds with complex constituents should not in a greater number take part in the variability of spelling.

It is clear that my observations on type frequencies, structure and spelling of English compounds need to be confirmed in a much larger corpus. Objections could be raised that these observations merely reflect the specificities of spelling of compounds in the LDCE, but

I believe that the correlation with the structure of compounds points to their more general holding.

Notes

[1] I express here my deep gratitude to an anonymous reviewer on the first version of this paper for the detailed comments. A version of this paper was presented at SDAŠ conference in Maribor in October 2008.

[2] Lieber notes that "NN compounds are very productive. Nearly any two nouns can be concatenated to form a new compound – for example, *armadillo dog,* which can in turn be compounded with another noun to form a still longer compound, *armadillo dog symposium.*"

[3] The linguists we mention here probably had in mind compounds written with an interword space. These are very conspicuous in written English, especially in the press. The compounds written together seem not to occur often in English (cf. Juhasz et al. 2003: 228).

[4] The situation is, however, changing fast, especially in psycholinguistic literature. One exception known to me is the study of Inhoff et al. (2000), who examined not only tripartite German compounds, but also compounds with constituents derived by adding the suffix *-ung*; they found that the subjects more rapidly recognised morpheme boundaries marked by a space than by the suffix *-ung.*

[5] Selkirk (1982) and Bauer (1983) only note which word classes can become constituents of nominal compounds. The productivity of noun – noun compounding has been discussed in more detail in Kiefer (2001). Štekauer (2005) contains a thorough discussion of meaning predictability, a question closely connected to productivity.

[6] Miller (2006) notes that spontaneous speech is "subject to the limitations of short-term memory." In such speech "phrases contain fewer words and clauses contain fewer phrases" than in planned writing. He also notes that a similar pattern holds for compound nouns. On the other hand, the LDCE (2000) itself is strongly based on spoken language. One of the consequences of such an orientation is that polymorphous compounds are not at all numerous in our corpus.

[7] It is important to note the difference between compound types and tokens (Lyons 1977: 6). Here we are mainly concerned with compound types.

[8] I accept here the terminology of Quirk et al. (1985) for the different ways of spelling English compounds.

[9]Nouns like *arts, sales* and *studies* for which there are the nouns *art, sale* and *study* with similar meaning have been tagged as morphologically complex.

[10]If an English speaker recognises the word *day* in *Wednesday* as a part of this noun, this seems to be the only option left. This decomposition can of course take place only in dialects of English in which the segment *day* has kept its pronunciation /deI/. Booij (2002) dubbed the words with bound roots as 'formally complex' since their prosody or some other property may indicate that they are complex. In the above examples the clusters /nzd/ in *Wednesday* and /sb/ in *iceberg* are improbable phoneme combinations for monomorphemic words in English. Taft (2004) argued that bound roots also have some kind of mental representation – the words which have frequent bound roots are easier to recognize than words which do not. Here, as in linguistics in general, we find that the boundaries between categories can be fuzzy, but the dividing line has to be drawn at some point – we have to set

boundaries somewhere in order to be able to start the analyses. Through analysis, initial categorization can be tested, and possibly corrected. For a similar view see Aarts (2004).

[11] Assumedly, in this case the 'direct route' has a good chance to win according to the widely accepted race model of lexical access (Frauenfleder and Schreuder 1992, Baayen 1992). The outcome of the race usually also depends on the frequency of constituents (cf. Hay 2001, de Jong et al. 2002, Juhasz et al. 2003).

[12] In Rakić (2008) I tried to show that the phonotactics of the morpheme boundary also plays a role in the spelling of compounds.

[13]Similarly the noun *vacuum* is classified as a derivative with a suffix *-um* since it is paradigmatically connected with the adjective *vacuous* and the noun *vacuity* which are derived from the suffixes *-ous* and *-ity* from the same base.

[14] The noun *compartment* is an example of prefixal-suffixal derivation in English because it is not possible to parse it into *compart +ment* or into *com+partment*. Such derivations are known in other languages (cf. Booij 2002, Клајн 2003), but are rarely found in English. We classified the noun *compartment* here as a prefixal derivation simply because there are so few prefixal derivatives.

[15] The difference between phrases and compounds in English is still a subject of controversies and discussions. We apply here the proposition that every combination noun +noun may be a compound if it has a definite denotation (cf. Plag 2006, Bauer 1998, Munat 2002).

[16] In the compounds *five-star general* and *four-letter word* the first components are not prototypical noun phrases because they do not obey number agreement. These compounds are probably best interpreted as special constructions (Booij 2005).

[17] In the compound *fly-drive holiday,* the first constituent *fly-drive* is a verbal compound.

[18] See also Aarts (2004), who argued that hybrid categories are disfavoured by language users because they are hard to process mentally.

[19] 598 compounds now make 100%.

[20]Hay (2003) convincingly argues that improbable phonotactics significantly facilitates a decomposition route. "If an affixed word possesses no such properties, however, no boundary will be assigned." Another important factor is lexical frequency, especially the proportion of lexical frequency of the derived word to lexical frequency of its base. Let us note that in (4) the frequency of the verb *chat* and the noun *chatter* is equal in Francis and Kučera (1982).

[21]According to the terminology of Aarts (2004), these words are complex words which converge to the category of simplex words. Hay and Baayen (2005) would rather take these words for another case showing that morphological categories are inherently gradient.

[22] The same holds for the compounds *Brussels carpet, Brussels lace* and *Brussels sprout,* although *Brussels* is singular. In Rakić (2008) I have argued that phonotactic transition is an important factor influencing the spelling of English compounds.

[23] We also see the fallout of *-s* in the case of some nouns ending in *-ics.* The words *economic, optic, athletic* are tagged in the LDCE as attributive adjectives, and a number of others do not bear such a tag because they can sporadically be applied predicatively, as can be checked in the BNC (e.g. *acoustic, gymnastic, genetic, phonetic, semantic*).

[24] Aarts (2004) rightly warned about accepting the categorical distinctions of traditional grammar because closer scrutiny may detect that the criterial properties are not precisely defined. On the basis of some characteristic cases, he argues that defining properties should be strictly bound to those 'morphosyntactic in nature'. As Hay (2003) has shown, the situation is more complicated because words may be morphologically complex in different degrees, depending on the frequency and phonotactic transitions of their parts. We, however, have to start with some categorisations, and the most natural ones are those which are widely accepted.

[25] Taft (1988: 662) even observed that lexical frequency affects the recognition of words with bound stems. He gives the example of the verb *deploy,* which has the same frequency of occurrence as *deflate,* but is recognised more rapidly than *deflate.* In Taft's interpretation this follows from the fact that "the stem *ploy* is more common as a stem than is *flate*, since *employ* is more common word than is *inflate*". Similar frequency effects of bound stems in derived words were observed by Bradley (1981), but only with the transparent, productive affixes, *-ness, -ment* and *-er*; no such effect was obtained with the opaque, nonproductive suffix *-ion*.

[26] This is also the thesis of Taft (2004), who noted that the later stage of recombining the parts of a complex word may "counterbalance the advantage of easier access to the higher frequency stem".

[27]The adjusted frequency is a frequency calculated on the basis of actual frequencies. It is intended to correct the unequal distribution of words in different genres of the corpus. In (9), for nouns *comb* and *wear* only actual frequencies are given; the adjusted frequency for these words is lower than 5.

[28] The verb *wear* is included in Francis and Kučera's list, but not the noun.

[29] It is not by chance that some of the second constituents in (2) have been tagged as 'combining forms'– these lexical items are high-frequency words, they are easily parsable since they begin with consonants and, importantly, they massively appear as second constituents in compounds with somewhat reduced meaning. In our corpus, *man* appears in 125 compounds, *woman* in 33, *work* in 42, *way* in 27, but *person* in only 8 compounds *(anchorperson, catperson, chairperson, cityperson, nightperson, morningperson, salesperson, weatherperson*). The last examples show the effect of a newly introduced convention according to which *person,* as a neutral formative, should replace the older *man* and *woman* – so far the convention has not affected a large number of compounds. Some other ending constituents in (2) like *girl, maid* or *folk* can be considered to be halfway towards gaining the status of combining forms since they take part in compound-like combinations with reduced meaning, and considerably more often in the second than in the first constituent. But in this area it is impossible to make predictions with any certainty because of the fuzzy boundaries which may have existed for a long time or even indefinitely (Hopper and Traugott 2003). The variability of the lists of combining forms cited in the literature suggests that one should proceed with caution in dealing with combining forms. They are paradigmatic examples of the morphological items exhibiting gradient properties (Hay and Baayen 2005).

[30] At least this is the way such nouns are treated in lexical phonology and morphology (Kiparsky 1982).

[31] For a similar view see Bauer (2001: 123). If the dual-route model of speech recognition is assumed, as is often the case nowadays, the understanding of recursive compounds does not seem to be an easy operation, especially if the message is encoded in speech sounds.

References

AARTS, Bas. 2004. Modelling linguistic gradience. *Studies in Lamguage* 28 (1), pp. 1-49.

ANDREWS, Sally, MILLER, Brett and RAYNER, Keith. 2004. Eye movements and morphological segmentation of compound words: There is a mouse in mousetrap. *European Journal of Cognitive Psychology* 16, pp. 285-311.

ALLEN**,** Margaret. 1978. *Morphological Investigations*. Doctoral dissertation, University of Connecticat, Sttors.

BAAYEN, Harald. (1992) On frequency, transparency, and productivity. In: G.BOOIJ and J. van MARLE (eds), *Yearbook of Morphology* 1992. Dordrecht: Kluwer Academic Publishers, pp. 181–208.

BAUER, Laurie. 1983. *English Word-Formation.* Cambridge: Cambridge University Press.

BAUER, Laurie. 1998. When is a sequence of noun + noun a compound in English? *English Language and Linguistics* 2, pp. 65-86.

BAUER, Laurie. 2001. *Morphological Productivity.* Cambridge: Cambridge University Press.

BAUER, Laurie and HUDDLESTON, Rodney. 2002. Lexical word-formation. In: R.

HUDDLESTON and G. K.PULLUM (eds.): *The Cambridge Grammar of the English Language*. Cambridge, New York, Port Melbourne, Madrid & Cape Town: Cambridge University Press. pp. 1621-1722.

BOOIJ, Geert. 1996. Autonomous morphology and paradigmatic relations, In: *Yearbook of morphology* 1996, 35-53.

BOOIJ, Geert. 2002. *Dutch Morphology.* Oxford: Oxford University Press.

BOOIJ, Geert. 2005. Compounding and derivation: evidence for Construction Morphology. In: Wolfgang U. DRESSLER, Franz RAINER, Dieter KASTOVSKY and Oskar PFEIFFER (eds.), *Morphology and its Demarcations*. Amsterdam: John Benjamins. pp.109-132.

BRADLEY, Dianne. 1981. Lexical representation of derivational relation. In: M.ARONOFF and M.L. KEAN (eds.), *Juncture.* Saratoga, Ca.: Anma Libri. pp.37-57.

COED. 2004. *Concise Oxford English Dictionary*, Eleventh edition. Oxford: Oxford University Press.

COLLINS COBUILD ENGLISH GUIDES 2. Word formation 1991. London: Harper Colins Publishers.

CUTTLER, Anne, HAWKINS, John and GILLIGAN, Gary. 1985. The suffic prefference: a processing information. *Linguistics* 23, pp. 723-758.

DE JONG, Nivja, FELDMAN, Laurie, SCHREUDER, Robert, PASTIZZO, Matthew and BAAYEN, Harald. 2002. The processing and representation of Dutch and English compounds: peripheral morphological and central orthographic effects, *Brain and Language* 81, pp.555-567.

DOWNING, Pamela. 1977. On the creation and use of English compound nouns. In: *Language* 53: pp. 810-842.

FABB, Nigel. 1998. Compounding. In: Andrew SPENCER and Arnold ZWICKY (eds.), *The Handbook of Morphology*. Oxford: Basil Blackwell, pp. 66-83.

FRAUENFELDER, Ulrich and SCHREUDER, Robert. 1992. Constraining Psycholinguistic model of Morphological Processing and Representation: the Role of Productivity. In: G. BOOIJ and J.van MARLE (eds.), *Yearbook of Morphology* 1991.Kluwer Academic Publishers, Dordrecht, pp. 165-185.

FRANCIS, W. Nelson and KUČERA, Henry. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston: Houghton Mifflin Company.

FRISSON, Steven, NISWANDER-KLEMENT, Elizabeth and POLATSEK, Aleksander. 2008. The role of semantic transparency in the processing of English compound words. *British journal of psychology* 99, pp. 87-107.

HAY, Jennifer. 2001. Lexical Frequency in morphology: Is everything relative? *Linguistics* 36 (6), 1040-1071.

HAY, Jennifer. 2003. *Causes and Consequences of Word Structure.* New York & London: Routledge.

HAY, Jennifer and BAAYEN, Harald. 2005. Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Sciences* 9 (7), pp. 342-348.

HAYES, Jenny, MURPHY, Victoria, DAVEY, Neil, SMITH, Pamela and PETERS, Lorna. 2002. The /s/ morpheme and the compounding phenomenon in English. In: W.D.GRAY & C.D. SCHUNN (eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society.* Erlbaum, Mahwah, NJ.

HOOPER, Paul and TRAUGOTT, Elizabeth Clos. 2003. *Grammaticalization.* Second edition. Cambridge: Cambridge University Press.

HUIJER, Thalia.and KROTT, Andrea. 2004. Access and representation of open and concatenated English compound words: the role of the constituent family revisited. The 10th Annual Conference on Architectures and Mechanisms of Language Processing, September, 16-18, 2004. Laboratoire Parole et Langage, CNRS,Université d'Aix.

JESPERSEN, Otto. 1942. *A Modern English Grammar on Historical Principles,* vol. VI: Morphology. Reprinted 1961. London: Allen & Unwinn.

INHOFF, Albrecht, RADACH, Ralph and HELLER, Dieter. 2000. Complex Compounds in German: Interword Spaces Facilitate Segmentation but hinder assignment of Meaning. *Journal of Memory and*

Language 42, pp. 23-50.

INHOFF, Albrecht, STARR, Matthew, SOLOMON, Matthew, and PLACKE, Lars. 2008. Eye movements during the reading of compound words and the influence of lexeme meaning. *Memory and Cognition* 36 (3), pp. 675-687.

JUHASZ, Barbara, STARR, Matthew, INHOFF, Albrecht and PLACKE, Lars. 2003. The effect of morphology on the processing of compound words: Evidence from naming, lexical decision and eye fixation. *British Journal of Psychology* 94, pp. 223-244.

JUHASZ Barbara, INHOFF, Albrecht, RAYNER, Keith. 2005. The role of interword spaces in the processing English compound words. *Language and cognitive processes* 20**,** pp. 291-316.

KATAMBA, Francis 1994. *English Words.* London and New York: Routledge.

KIEFER, Ferenc. 2001. Productivity and compounding. In: C. SCHANER-WOLLES, J. RENNISON and F. NEUBARTH (eds.), *Naturally! Linguistic studies in honour of Wolfgang Ulrich Dressler presented on the occasion of his 60th birthday*. Torino: Rosenberg and Sellier, pp. 225-231.

KIPARSKY, Paul. 1982. Lexical Morphology and Phonology. In: I.S. YUNG (ed.), *Linguistics in the Morning Calm.* Seoul: Hanshin, pp. 1-92.

КЛАЈН, Игор. 2002. *Творба речи у српскоме језику. Први део, слагање и префиксација.* Београд: Одбор за стандардизацију српског језика.

LDCE 2000. *Longman Dictionary of Contemporary English*, Third Edition. Edingurgh Gate, Harlow, Essex: Longman.

LEECH, Geoffrey, RAYSON, Paul and WILSON, Andrew. 2001. *Word Frequencies in Written and Spoken English: based on the British National Corpus.* London: Longman.

LEVINSON, C. Stephen. 1983. *Pragmatics.* Cambridge: Cambridge University Press.

LIBBEN, Gary, GIBSON, Martha, YOON, Yeo Bom and SANDRA, Dominiek. 2003. Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language*, 84 (1), pp. 50-64

LIEBER, Rochelle. 1992.Compounding in English. In: *Rivista di Linguistica* 4 (1), pp. 79-96

LYONS, John 1977. *Semantics,* vol. 1. Cambridge: Cambridge University Press.

MARCHAND, Hans. 1969. *The Categories and Types of Present-day English Word-Formation*, Second edition. München: Beck.

MACWHINNEY, Brian and SNOW, Catherine.1985. The Child Languiage Data Exchange System. In: *Journal of Child Language* 12, pp. 271-296.

MILLER, Jim. 2006. Spoken and Written English. In: B.AARTS and A.MACMAHON (eds.), *The Handbook of English Linguistics.* Oxford: Blackwell Publishing, pp. 670-691.

NORRIS, Dennis and CUTLER, Anne.1985. Juncture Detection. *Linguistics* 23, pp.680-705.
PLAG, Ingo. 2003. *Word Formation in English*. Cambridge: Cambridge University Press.

PLAG, Ingo. 2006. The variability of compound stress in English: structural, semantic and analogical factors. *English Language and Linguistics* 10 (1), pp.143-172.

PRĆIĆ, Tvrtko. 2005. Prefixes vs. initial combining forms in English: a lexicographic perspective. *Internationa Journal of Lexicography* 18 (3), pp. 313- 335.

QUIRK, Randolf, GREENBAUM, Sidney, LEECH, Geoffrey and SVARTVIK, Jan (eds.).1985. A *Comprehensive Grammar of the English language*. London & New York: Longman.

RAKIĆ, Stanimir. 2008. Some further observations on the Spelling of English compounds. Presented at SDAS Conference in Maribor (appears in *ELOPE* 6).

SELKIRK, Elisabeth. 1982. *The Syntax of Words*. Cambridge. Ma.: MIT Press.

SPROAT, Robert. 1985. *On Deriving the Lexicon.* Ph.D. thesis. Massachusets Institute of Technology.

ŠTEKAUER, Pavol 2001. Fundamental Principles of an Onomasiological Theory of English Word-Formation. *Onomasiology* 2. pp. 1-42. Online [www.onomasiology.de].

ŠTEKAUER, Pavol. 2005. Meaning Predictability in Word Formation: Novel, Context-free Naming Units. Amsterdam: John Benjamins.

TAFT, Marcus. 1988. A morphological-decomposition model of lexical representation. *Linguistics* 26, pp. 657-667.

TAFT, Marcus. 2004. Mophological decomposition and the reverse base freqency effects. *The quarterly journal of experimental psychology* 57A, pp. 745-765.

TAFT, Marcus and FORSTER, Kennrth, I. 1975. Lexical storage and retrieval of prefixed words, *Journal of Verbal Learning and Verbal Behavior* 14, pp. 638-647.

WIESE, Richard. 1996. Phrasal compounds and the theory of word syntax. *Linguistic Inquiry* 27 (1), 1996, pp. 183-193.

WIESE, Richard. 2000. The Structure of the German vocabulary: edge marking of categories and functional considerations. *Linguistics* 36 (6), pp. 95-115.

Stanimir Rakić
Blv. Arsenija Čarnojevića 37/26
11070 Belgrade
Serbia.
starakic@yahoo.com

In *SKASE Journal of Theoretical Linguistics* [online]. 2009, vol. 6, no. 1 [cit. 2009-07-06]. Available on web page <http://www.skase.sk/Volumes/JTL13/pdf_doc/04.pdf>. ISSN 1339-782X.