The language of A Clockwork Orange: A corpus stylistic approach to Nadsat

Vincent, B & Clarke, J

The language of *A Clockwork Orange*: A corpus stylistic approach to Nadsat

Benet Vincent and Jim Clarke
Coventry University, UK

Abstract
The 1962 dystopian novella *A Clockwork Orange* achieved global cultural resonance when it was adapted for the cinema by Stanley Kubrick in 1971. However, its author Anthony Burgess insisted that the novel's innovative element was the introduction of 'Nadsat', an art language he created for his protagonist Alex and his violent gang of *droogs*. This constructed anti-language has achieved a cultural currency and become the subject of considerable academic attention over a 50-year period, but to date no study has attempted a systematic analysis of its resources and distribution. Rather, a number of studies have attempted to investigate the effects of Nadsat, especially in terms of the author's claim that learning it functioned as a form of 'brainwashing' embedded within the text.

This paper uses corpus methods to help isolate, quantify and categorise the distinctive lexicogrammatical features of this art language and investigate how Burgess introduces a new, mainly Russian-based lexicon to readers. In doing so, it clarifies the existing confusion over what Nadsat is, and also provides a roadmap for future studies into the construction, function and translatability of the created linguistic component of the novel.

1. Introduction

The English polymath Anthony Burgess is best known for his 1962 novella *A Clockwork Orange* (*ACO*), later adapted for the cinema by Stanley Kubrick. It relates the story of Alex, the charismatic protagonist and narrator, who leads a futuristic gang of teen thugs on a spree of violence and sexual assault culminating in his arrest, imprisonment and attempted brainwashing by the authorities. The morality tale which ensues explores the limitations of human free will and its interaction with good and evil. It has been translated into more than 30 languages.

The novella is notable for the constructed artistic language, or 'art language',[1] 'Nadsat', which is spoken by the protagonist–narrator Alex and his *droogs*,[2] or friends, and gives a distinctive flavour to the narrative voice through its incorporation of Russian and other components; *-nadsat*, the suffix used for the numbers 11–19 in Russian, is effectively the Russian counterpart of *-teen*. Burgess was quite protective of, and had clear ideas about, the function of his invented language. One of his reasons for persistently disowning *ACO* following the release of the controversial Kubrick adaptation was that the film failed to foreground the linguistic medium he had carefully constructed (Burgess, 1990).

As a keen linguist, Burgess was in fact well placed to approach the issue of inventing an art language such as Nadsat. A lifelong philologist, he produced linguistics textbooks such as *Language Made Plain* (1964) and *A Mouthful of Air* (1992), as well as other art languages such as 'Ulam', the reconstruction of proto-Indo-European created for Jean-Jacques Annaud's (1981) film *Quest for Fire*. This background linguistic knowledge may account for the success of Nadsat and the attention it has attracted. Elements of Nadsat, such as the word *droog* itself, have since passed into popular culture (Jackson, 1991) and many online glossaries and commentaries on Nadsat exist. The distinctive use of language in the novel, in particular Nadsat, has also attracted considerable attention from literary studies (e.g. Carson, 1976; Evans, 1971; Fowler, 1979; Jackson, 2011; Lennon, 2010), translation studies (Maher, 2010, 2011) and even studies of vocabulary acquisition (Pitts et al., 1989; Saragi et al., 1978). However, despite offering analyses of Nadsat, most, if not all, of these studies largely rely for their description of this 'art anti-language' (Fowler, 1979) on what Lennon (2010: 101) terms 'pseudo-scholarly supplements to the novel', chiefly Stanley Hyman's afterword and glossary, which were appended to the 1963 Norton US edition of *ACO*. They typically do so, however, without referring to

Hyman's admission therein that the glossary was not only unauthorised but also partly guesswork (Hyman, 1963).

While reliance on unauthorised lists of Nadsat does not in itself invalidate earlier work, the lack of clarity surrounding its characterisation and realisation creates difficulties of comparison between studies; at least three different and conflicting glossaries (Biswell, 2012; Hyman, 1963; Rawlinson, 2011) as well as numerous categorisations with varying levels of precision exist.

As a result, despite Nadsat's significance in popular culture (Biswell, 2012), and its influence since its publication upon the development of constructed art languages within literature, for example Jeff Noon's *Vurt* (1993), there has been no comprehensive attempt to define and delimit Nadsat. The fuzziness that therefore remains is of particular concern to us as we are part of a project that seeks to investigate what happens to Nadsat when it is translated into other languages. The aim of this paper is to rectify that lack of definition and show how corpus approaches may be useful in investigating this art language.

2. Previous attempts to describe Nadsat

2.1 Burgess's perspective

In a lifelong attempt to define his literary achievements for readers (including critics, who he felt tended to misunderstand his fiction), Burgess frequently laid out his authorial intentions in articles, memoirs and other non-fiction works (e.g. Burgess, 1972, 1990). This need to explain himself particularly pertains to *ACO*, which Burgess (1990) argued had been misrepresented by both Stanley Kubrick's cinematic adaptation and the cultural reception of the movie version. However, Burgess's inconsistency and (perhaps deliberate) vagueness about what Nadsat is appears to have contributed to inconsistencies and vagueness in treatments of this art anti-language.

In a letter to Heinemann, the first publishers of *ACO*, Burgess claimed that '[i]t will take the reader no more than fifteen pages to master and revel in the expressive language of "nadsat"' (Biswell, personal communication, 2016).[3] It is clear Nadsat was designed to be intelligible to a reader within a short period of encountering it without the need for further help.[4] Burgess (1972, 1990) also said that his intention was for the reader to be in some sense 'brainwashed' in this process of acquiring the Russian elements of Nadsat. This focus on Nadsat as essentially Russian vocabulary mixed into English, however, oversimplifies what it is and stands in contradistinction to its characterisation within the book itself, where two psychologists examining Alex describe this 'dialect of the tribe' as composed of 'Odd bits of old rhyming slang,' ... 'A bit of gypsy talk, too. But most of the roots are Slav' (Burgess, 2012: 125). While we do not of course have to take this statement at face value, it gives an indication that Nadsat is composed of different elements. At the same time, the intention for Nadsat to be acquired in the process of reading the novel might explain why Burgess was unwilling to give many clues about its composition.

2.2 Nadsat in context: Issues of identification

The following extract, from the first page of *ACO*, is illustrative of how Nadsat is realised in the book.

> Our pockets were full of deng, so there was no real need from the point of view of crasting any more pretty polly to tolchock some old veck in an alley and viddy him swim in his blood while we counted the takings and divided by four, nor to do the ultra-violent on some shivering starry grey-haired ptitsa in a shop and go smecking off with the till's guts. But as they say, money isn't everything.[5] (Burgess, 2012: 7–8)

Several points relevant to the description and identification of Nadsat can be raised in relation to this extract. Firstly, there is no identification of particular items as being Nadsat, such as by the use of a different typeface. This means that – except for a few places where Nadsat items are glossed by Alex (see section 4.3) – the identification of Nadsat items will rely on reader interpretation. While for most readers the obviously foreign (mainly Russian) relexicalisations such as *deng* for "money/cash" will stand out, other potential candidates, such as *pretty polly* and *ultra-violent* are not so clear cut. That is, using Leech and Short's (2007) terms, while it may be possible to agree on which words are 'deviant' with respect to standard English and therefore candidates for Nadsat, different items may be 'prominent' (and therefore 'foregrounded') for different speakers depending on the speaker's background.

A second aspect of the above extract, one that Hyman (1963), Lennon (2010) and Maher (2010) have noted with regard to *ACO*, is that, in line with his aim of encouraging readers to acquire Nadsat

with ease (Burgess, 1972), Burgess takes trouble to present the novel lexis of Nadsat in contexts where a reader may make at least a reasonable stab at understanding it. Knowing, for example, that their pockets were 'full of' *deng* limits what *deng* might mean, particularly as it appears to be an uncountable noun in Nadsat (as *money* is in English); other references to ill-gotten gains (*the takings, the till's guts, money*) could also be seen as helping to clarify what is meant here. It is also noteworthy that at least two other words in the extract – *pockets* and *blood* – have Nadsat equivalents (*carmans* and *krovvy*, respectively) that are not used here. That is, while there appears to be a high concentration of Nadsat items in this extract (by our reckoning, nearly double that in the book as a whole, without even considering *pretty polly* and *ultra-violent*[6]), Burgess has arguably avoided using other possibilities so that the passage may remain comprehensible.

2.3 Nadsat as lexicon: The glossary approach

Despite the fact that this was contrary to Burgess's explicit wishes, the relatively high concentration of potentially unfamiliar words in extracts like the one in section 2.2 led the literary critic Stanley Hyman to create a 241-item glossary which accompanied the first US edition of *ACO* in 1963. As a work unauthorised by the author, Hyman had no help from Burgess and, as noted in section 2.1, was reliant to some extent on guesswork. Unfortunately, he also did not know Russian, but instead had to rely on a colleague and some correspondents, which introduced some errors.[7] Moreover, Hyman's lack of knowledge of British-English slang and omission to check relevant sources such as the *Oxford English Dictionary* (*OED*) led him to include or miscategorise English slang words that he was apparently unfamiliar with, such as *sod, snuff it* and *rozz*. Finally, some items, such as *eggiweg* and *warble* appear in the 'glossary' not because they need glossing but apparently because Hyman classes them as Nadsat; no means of deciding what Nadsat is or is not is suggested, however.

Alternative glossaries are also provided by Rawlinson (2011) and Biswell (2012)[8] in other editions of *ACO*. These, however, do not make the picture any clearer. Rawlinson only includes items he interprets as deriving from Russian but seems to be less comprehensive than Hyman (177 items compared with 187 that Hyman classes as having Russian origins), while Biswell (2012) includes 202 items in the 'expanded glossary', although this is considerably smaller than Hyman's 241-item list, which is divided into items of Russian origin and others.

In short, the glossaries provided with different editions, apart from being contrary to Burgess's wishes, are to some extent contradictory and not entirely satisfactory, containing a number of apparent errors. Criteria are not provided for deciding what Nadsat is and it is thus hard to understand how certain items were included. If these glossaries were only used by readers of *ACO* to deal with comprehension errors, such differences would be of minor interest. However, as noted in section 1, a number of studies have used Hyman's list as a starting point for their research without apparently noting the potential issues its rather haphazard compilation might create. To take two examples, Pitts et al. (1989: 271) refer to Nadsat as '241 slang words of Russian origin', which in fact misrepresents Hyman's glossary, while Fowler (1979: 268) refers to Nadsat by saying that 'Burgess provides a glossary of around 250 words'.

2.4 Categorisations of Nadsat

The lack of clarity relating to glossaries of Nadsat is also seen when it comes to categorisations of Nadsat items. This is a commonly undertaken procedure that seeks to divide terms according to their derivations and/or formal features; it tends to separate items that are likely to be incomprehensible to the average reader (i.e. Russian-derived words) and other items which, while not necessarily presenting problems of comprehension, are distinctive to Alex and the *droogs*' speech.

As noted in section 1, the categorisation approach to Nadsat is suggested in the novel itself by reference to its composition as 'Slav' mixed with 'bits of old rhyming slang', and 'A bit of gypsy talk' (Burgess, 2012: 125). We have already seen examples of 'Slav' (i.e. Russian) and an example of rhyming slang (*pretty polly* rhymes with, and therefore refers to, 'lolly', which is itself a slang term for money) in the analysis of the extract in section 2.2. It is not clear to what extent, however, Burgess wants us to take this analysis seriously. Few, if any, commentators have been able to find evidence of Romani items in the Nadsat lexicon (Jackson, 2011) and a number of other classes of items have been identified, most of which can be traced to Hyman's (1963) afterword.

The first of these Hyman refers to as 'schoolboy transformations'; this category, elsewhere labelled 'reduplications' (Jackson, 2011) or 'schoolboy infantilism' (Lennon, 2010), includes items such as *appy polly loggies*, *baddiwad* and *eggiweg*. These items are relatively easy to recognise, since they are included on the basis of clear formal characteristics alluded to in Jackson's (2011) label and apparent in

the three examples shown here: *i, ie* or *y* is added to each syllable of polysyllabic words (*apologies*) while monosyllabic words either repeat the syllable or have the end of the syllable repeated after (*i)w*.

A second category of items noted by Hyman is 'amputations', or shortened words; Evans (1971) calls these 'clipped forms' and Lennon (2010) refers to 'truncation'. Hyman gives a few examples, such as *guff* ("guffaw") and *sinny* ("cinema") but does not attempt an exhaustive list.

Hyman's last category is 'portmanteau words', that is, new words created using features of two existing words and including elements of the meanings of the source words. Examples provided include *chumble*, which he claims is a combination of 'chatter' and 'mumble' and *skriking* ('scratching' + 'striking'). Despite both Lennon's (2010) and Jackson's (2011) endorsements of Hyman's points here, this is a problematic category. None of the words listed by Hyman as portmanteau words stand up to investigation: to take two examples, *chumble* is listed in the *OED* ("gnaw, nibble, peck"), while *skriking* seems more likely to have been taken from the Russian 'vskriknut' meaning "(to) scream".

A further aspect of Nadsat noted in a number of other studies (Carson, 1976; Maher, 2010, 2011) is the use of archaic language such as now obsolete pronouns (in standard English, at least) *thou, thee, thy* as well as some archaic lexis. It is interesting to note that at least one of these archaic words, *darkmans* ("night"), is also mentioned by Halliday (1976) as part of Elizabethan thieves' cant in his discussion of anti-language. As with other categories of non-Russian Nadsat items, the extent of archaisms in *ACO* has not previously been investigated.

In sum, categorisations of Nadsat tend to be fairly speculative and make no claims to comprehensiveness. In no case are guidelines presented for deciding whether a particular item is or is not Nadsat, and a number of later scholars follow Hyman's example of relying on speculation and intuition, with the result that a number of erroneous interpretations and categorisations can be seen. Our aim in the rest of this paper is to outline the means by which we sought to overcome these perceived shortcomings. The approach taken is essentially a corpus stylistic one, that is, 'the application of corpus methods to the analysis of literary texts by relating linguistic description with literary appreciation' (Mahlberg, 2013: 5).

3. Methodology: Identifying Nadsat

In order to investigate Nadsat in the novel, we treated *ACO* as a corpus (the electronic version of the text was obtained with the help of the International Anthony Burgess Foundation). The edition used was the 2012 'Restored Edition', edited by Andrew Biswell (Burgess, 2012), which includes the 21st chapter of the book, originally omitted from early American editions. The chapters were separated into different files to enable us to track the extent and distribution of Nadsat use across the book.

There seems general consensus that Nadsat is an anti-language (Fowler, 1979; Kohn, 2008), a term used by Halliday (1976) to refer to a language variety used by a group that sees itself as in opposition to the values of a society. This inherent contrast to linguistic norms suggests an obvious approach to identify candidates for inclusion – keyword analysis. Keyword analysis, an approach pioneered by Scott (1997), has increasingly been used in corpus stylistics as a means of approaching deviation and thereby foregrounding (Leech and Short, 2007; Mukařovsky, 1958), that is, aesthetically motivated linguistic deviation from a particular norm (Johnson, 2016). The keyword approach is particularly suited to investigating foregrounding because it provides the analyst with a list of wordforms that are found significantly more often in the target text or corpus than in a reference corpus (O'Halloran, 2007; Scott, 1997; Stubbs, 2005).

The most commonly used way of calculating the keywords of a target text or corpus measured against a reference corpus uses Dunning's (1993) log likelihood statistic. Although this method has seen wide usage in corpus linguistics and corpus stylistics, Kilgarriff (2009), Gabrielatos and Marchi (2011) and Rayson (2013) have raised some questions regarding its operationalisation and theoretical bases. These studies point out that a keyword list calculated in this manner provides a confidence measure rather than an effect size measure. Items are included for which there is sufficient evidence of greater frequency in the target text without an indication of the *extent* to which such items are more frequent[9] (Gabrielatos and Marchi, 2011). However, since, on this basis, infrequently occurring Nadsat items in *ACO* may be excluded due to lack of evidence, this means of extracting keywords was clearly inappropriate to our task.

For this reason, we chose instead the 'simple maths' method operationalised in Sketch Engine[10] and explained in Kilgarriff (2009, 2012). This method – which essentially calculates the ratio of normalised frequencies between target and reference corpus, the 'Score' values in Table 1 – has the added advantage that one can specify by changing the 'parameter' (a figure added to the calculation to avoid dividing by 0) whether one wants to retrieve relatively more frequent or relatively less frequent items

(Kilgarriff, 2009). In our case, the corpus used for the purpose of identifying keywords was the largest available on Sketch Engine, the 19-billion-word English Ten Ten 2013 corpus[11]; no minimum frequency was specified and the parameter was set at '1' (the default, which favours less frequent items; see Kilgarriff (2009)). This method resulted in a keyword list of 4554 items, the ten top ranking of which are shown in Table 1 for illustrative purposes. It is clear that, with the exception of 'Dim', the name of one of the characters, these are all Nadsat items.

This list of keywords was a starting point for further analysis; it was thus important to use procedures to decide whether or not each item might be counted as Nadsat, with

Table 1. The ten top ranking keywords in *A Clockwork Orange*.

	word	freq (in *ACO*)	freq /mill	freq_ref	freq_ref /mill	Score
1	veck	144	2065.0	11	0.0	2065.0
2	viddy	132	1892.9	96	0.0	1885.9
3	horrorshow	107	1534.4	240	0.0	1519.3
4	malenky	99	1419.7	11	0.0	1420.0
5	Dim	130	1864.2	14495	0.6	1138.9
6	viddied	76	1089.8	23	0.0	1089.7
7	goloss	65	932.1	3	0.0	933.0
8	glazzies	65	932.1	7	0.0	932.8
9	gulliver	65	932.1	192	0.0	925.3
10	litso	64	917.8	9	0.0	918.4

Key:
ref = reference corpus, i.e. English Ten Ten Web Corpus 2013.
/mill = normalised frequency per million words.

the underlying principle being that such items must deviate from standard English. As discussed in Section 2, there are a number of ways in which this deviation has already been shown and which we were able to use as a guide.

The main procedure involved identifying 'deviant' forms, or items that differ formally from standard English. As noted above, the principal ways in which these items may differ are by: being (based on) words from other languages, mainly Russian; being truncated forms of English words; using archaic or obsolete words and morphology (e.g. *thou knowest*; *mounch*, an archaic spelling of munch); showing other creative morphology, that is, employing novel word formation processes on familiar words, such as the reduplication already mentioned ('bad' becomes *baddiwad*) and unusual suffixation to make new parts of speech (the adjective 'gorgeous' becomes the noun *gorgeosity*).

Once a candidate item was identified, its use in the book was checked to ensure that it occurs predominantly in the language of Alex and the *droogs*; occasionally some Nadsat items crop up in the language of other characters, e.g. fellow prisoners or other young people. This checking procedure revealed that some items, e.g. *worldcast*, are also explicitly glossed as belonging to others, typically using a variant of *what they called*[12] (e.g. 'Tonight was what they called a worldcast' (Burgess, 2012: 23)). Such items can be contrasted with those glossed by the narrator as belonging to his speech using expressions like *we called them*: 'shoulders ("pletchoes", we called them)' (Burgess, 2012: 8). Finally, where necessary, we consulted reference works, mainly the *OED* and Partridge's *Dictionary of Slang and Unconventional English* (Partridge, 1937; Partridge and Beale, 1984), a source that Burgess is known to have used (Biswell, 2012).

This procedure does not, however, cover all possibilities. Other ways apart from morphological structure in which items may be foregrounded by deviating from standard English relate to their semantics or syntax (Leech and Short, 2007). Thus it was important to check whether items on keyword lists were used in unusual senses or functions. A clear case of deviant semantics can be seen in examples of rhyming slang, where the expression used is not intended to be taken at face value but instead is to be understood as referring to a rhyming word. For example, *pretty polly* conventionally refers to a parrot, but in *ACO* it instead refers to the rhyming word, *lolly*, a slang term for money (Partridge and Beale, 1984). Other examples of standard words with non-standard meanings include *lighter* ("woman": 'wrinkled old lighters') and *cancer*, which is used metonymically to mean "cigarette".

As far as syntax is concerned, the item *ultra-violent* provides an example of deviation from normal usage. In the extract reproduced in section 2, this word occurs in 'nor do the ultra-violent on some starry grey-haired ptitsa' (Burgess, 2012: 8). Perhaps the most salient aspect about the choice of *ultra-violent* here is that it is arguably ungrammatical since the usual noun form is 'violence'. However, two other

features of the Nadsat phrase it occurs in, DO *the (old) ultra-violent on someone*, seem at least unusual and to therefore contribute to the foregrounding of this expression: the choice of verb, which more usually would be PERPETRATE, INFLICT or COMMIT, and the use of the definite article, which is quite marked (compare 'inflict ultra-violence on someone').

On the basis of particular usage of already unusual items in the language of the *droogs* then, we argue that it is possible to identify Nadsat items from the keyword lists. This identification is greatly facilitated by corpus methods, chiefly concordance analysis (i.e. using empirical evidence found in the co-text of items of interest).

In following this procedure, a small number of words were identified which function both as Nadsat and in their normal meaning; two examples are *tree*, which can refer to the Russian word for three, and *sharp*, which is also used as a Nadsat word for woman. Frequencies of Nadsat items presented in Section 4 exclude instances where such items have their usual meanings. We also disregarded onomatopoeic items, e.g. *clop* in 'my heart was going clopclopclop'.

4. Results: Nadsat

This section will discuss the main categories of Nadsat identified in this study, before looking at their distribution in the book. We also discuss the phraseology of Nadsat and how it bears on the introduction and comprehension of new items.

4.1 Emerging categories of English–Nadsat

The main categories of Nadsat items identified using the procedures set out in Section 3 are shown in Table 2 together with the number of items in each category.[13] Unsurprisingly, the largest category of Nadsat is what we term 'core' Nadsat, that is, the essentially Russian-based relexicalisation of English which dominates the extract shown in Section 2.2. This category consists of 218 headwords and includes ten items either derived from other languages (e.g. *tass*, a French word meaning "cup") or of uncertain etymology (e.g. *shilarny*). The count is based on lemmatisation (e.g. *ittying, ittied* are all regularly formed inflections of *itty* – "go"); identical forms which realise different parts of speech (e.g. *krovvy* – "blood", "bleed") are also treated as one entry on the list. Different forms which are not formed by means of regular English grammatical inflections (e.g. *droog, droogie*[14]) are listed as different words, while compound nouns such as *cal-coloured* are not counted as separate words. Differences from counts included in published glossaries (see Section 2.3) are due to the fact that these lists are not generally consistent in such matters or do not include all forms.

The second category consistently noted in the literature on Nadsat is that of 'babytalk', including items such as *eggiweg* and *purplewurple*. Only ten of these items are found in the text, not including the related ('core' Nadsat) *malchickiwick*, which is derived in similar fashion from the Russian *malchick* ("boy").

Table 2. Categorisation of Nadsat.

Category	Number of words	Example items
Core Nadsat	218	*bolshy, cal, droog, itty, lighter, tashtook*
Archaisms	36	*ashake, canst, thou/thee/thy/thine, redding*
Babytalk	10	*eggiweg, purplewurple*
Rhyming slang	5	*luscious glory, pretty polly*
Truncations	21	*guff, hypo, sinny*
Compound words	46	*afterlunch, bruiseboy, in-grin, ultra-violent*
Creative morphology	20	*appetitish, crunk, syphilised, cancery*

An even smaller set which is also commonly mentioned is items of rhyming slang. Only five of these were identified: *pretty polly* ("lolly" – money), *luscious glory* ("upper storey", i.e. "hair"), *twenty-to-one* ("fun"), *hound-and-horny* ("corny")[15] and the less obvious *sharp* ("woman"), which seems to be derived from 'sharp and blunt', listed in Partridge and Beale (1984: 1044) as rhyming slang for "cunt".[16]

Alex's use of archaisms was also noted in Section 2.4. Most of these items do not usually feature in glossaries since they do not create issues of comprehension. We identify 36 such items including *ashake, canst, thou/thee/thy/thine* and *redding*. One particularly productive archaism is the use of -*wise*, both as a suffix (*thiswise*; *day-wise*) and on its own.

As noted in Section 2.4, a number of previous studies have mentioned individual truncated items such as *sinny* ("cinema") and *guff* ("guffaw"). We identified 20 such items; it is interesting to note that some core Nadsat items are formed in similar ways: *veck* ("person") is a shortened form of *chelloveck* and *chasso* ("guard") is a shortened form of the Russian word 'chasovoi'.

Another productive means of creating new words in Nadsat is compounding, an example of which, *ultra-violent*, was seen in the extract in Section 2.2. The 46 items in this category are included on the basis that they make a novel item with a meaning that is not necessarily predictable from their component parts (e.g. *lipmusic* – "(blow a) raspberry") and/or form an unpredicted part of speech (e.g. 'No school this afterlunch' (Burgess, 2012: 50)). This compounding of standard English words to make new items builds on a similar process that can also be seen with 'core' Nadsat items combining with English words (e.g. *counter-veck*, *krovvy-dripping*).

The final group of Nadsat items identified in our analysis is those which involve creative uses of already existing English words. This may be by adapting the ending of the word (*appetitish* for *appetizing*; *clowny* as an adjective from *clown*), using it as a new part of speech (*blueing* as a verb meaning "making blue"), or changing the spelling to add new associations (*syphilised* as an ironic replacement for *civilised*). At 20 items, the number of words in this group is quite small. As with several other categories, these are all processes that are found with Russian-derived words: *govoreet*, for example, is only a verb in Russian but in Nadsat it also functions as a noun ('a quiet govoreet' (Burgess, 2012: 50)); Russian words are also adapted in transliteration to create punning associations, for example *lewdies* (Russian *lyudi* – "people" + 'lewd') and *horrorshow* (Russian *khorosho* – "good/well") (Maher, 2010).

4.2 Distribution of Nadsat in ACO

While the categorisation of Nadsat outlined in Section 4.1 gives an indication of the lexical resources of this variety, it does not indicate their distribution across the book. Which categories, in short, are the most frequent? This is a question of some importance for those interested in studying Nadsat and its effects in more detail. Figure 1 shows the distributions in *ACO* of all inflections of items that have been identified in each category, normalised per 10,000 words. As can be seen, 'core' Nadsat dominates in terms of frequencies; overall 'core' Nadsat items are met on average around once every 17 words,[17] and this category is over 35 times as frequent as the next most frequent, truncated words, and nearly ten times more frequent than all 'other' types combined. These 'other' categories can thus be seen as less central to Nadsat overall.

It is also important to consider the distribution of Nadsat items by chapter in the book and how this matches the overall arc of the narrative. The resulting distributions, which are normalised to allow comparison across chapters of different lengths, are shown in Figure 2. This distribution raises several points of interest.

Firstly, it is clear that Part 1 Chapter 6 has the highest concentration of Nadsat items. This chapter is in fact one of the pivotal chapters of the book, since it relates Alex and the *droogs'* attack on an old lady in her home which eventually leads to her death and Alex's arrest and subsequent incarceration. The fact that this chapter involves mainly those characters who speak Nadsat explains why the concentration of these items is higher here. In contrast, the two chapters with the lowest concentrations of Nadsat items, Part 2, Chapter 3 and Part 3, Chapter 4, involve Alex's interactions with adult characters. In Part 2, Chapter 3 Alex is told by the prison governor that he is to have the aversion therapy that will 'cure' him of his violent tendencies; the treatment is explained in more detail by the doctors who will carry it out. Part 3 Chapter 4 relates Alex's second encounter with the writer F. Alexander (whom he earlier attacked and whose wife Alex and the *droogs* had earlier raped), who takes him in when Alex is wandering desperate and homeless along the road. The fact that Nadsat is still found in these chapters can be explained by the fact that Alex as narrator still uses it in his description of events, even if the dialogue between characters is closer to standard English.

Another point of interest regards the concentration of Nadsat in Chapter 1, where it is first introduced. We can see that 'core' Nadsat items, which represent the main challenges to comprehension, are comparatively infrequent (495 per 10,000 words, or around 15% lower than the figure for the book as a whole), while other Nadsat items, which do not typically create the same challenges, are more frequent (104 per 10,000 words, or 65% more frequent than in the book overall). This gives one indication that Chapter 1, where a high proportion of the Nadsat lexicon is introduced, features significant authorial effort to avoid oversaturation and ensure a degree of comprehensibility

while achieving the desired defamiliarisation effects (Fowler, 1979) by mixing in other elements attributable to the variety.

4.3 The phraseology of Nadsat

Considering Nadsat in terms of distributions of items through the book only provides part of the explanation of how it works. It is illustrative in this respect to consider some of the strategies that are used to introduce items and consider in what ways they help
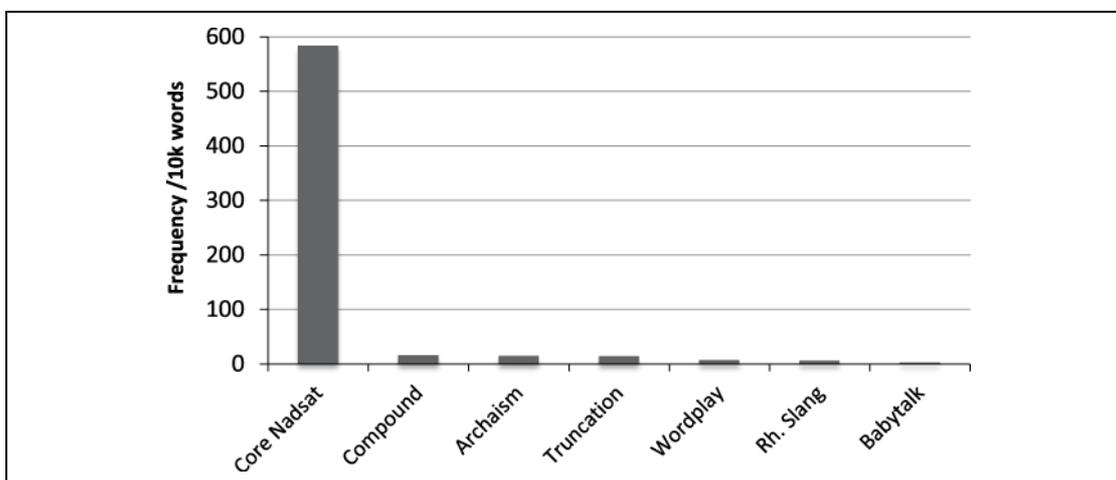

Figure 1. Normalised frequencies of all forms of items in the different categories of Nadsat.
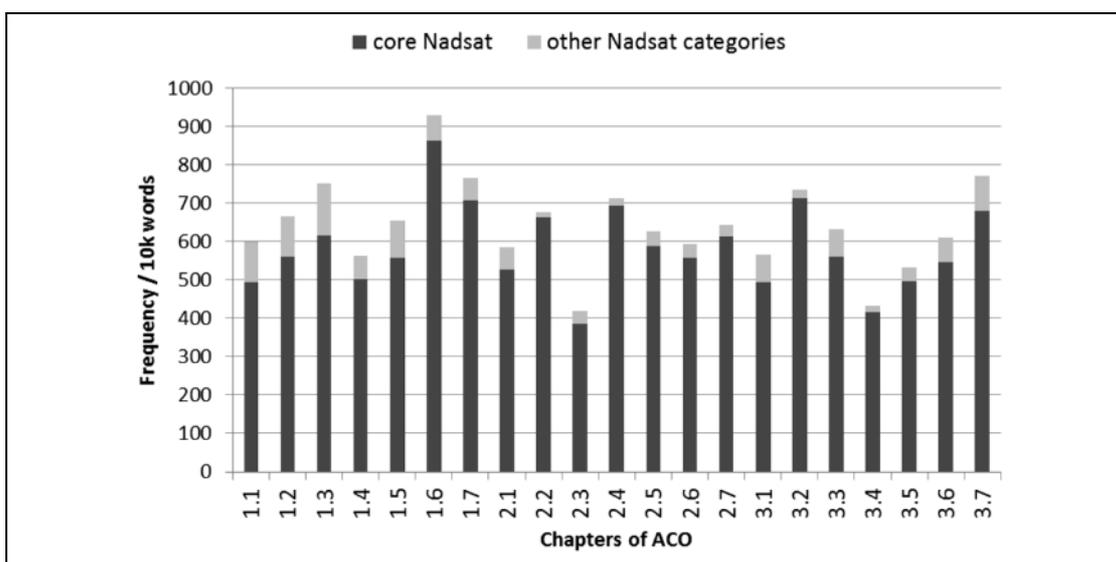
Figure 2. Distribution of Nadsat by chapter (1.1 = Part 1, Chapter 1).

the reader to acquire this language and become 'brainwashed'. The discussion below will therefore consider only 'core' Nadsat items. Investigating the phraseology of Nadsat items – i.e. how they are typically used (Sinclair, 1991) – also indicates some of the syntactic deviations of Nadsat compared to standard English, a feature pointed out but not pursued by Maher (2010).

A number of previous commentators have noted that some Nadsat items are glossed in the book itself; Fowler (1979) and Lennon (2010), indeed, claim they are 'often' glossed. In fact, a search of the glossing devices used in the book such as 'or' in *Bog or God*,[18] reveals that only 17 separate Nadsat items are glossed. Other means are therefore used to ensure that it is possible to understand Nadsat. These can be demonstrated by considering more frequent items, e.g. *veck* (144 instances) and *smecking* (30 instances), as well as those that occur once or twice, since the strategies used tend to differ accordingly.

As it is the most frequent Nadsat wordform in the book, there are plenty of chances for the reader to encounter *veck* ("man") in context. This may be one reason why the first encounter with *veck* – seen in the extract in Section 2 earlier (*to tolchock some old veck in an alley and viddy him swim in his blood*) – is not glossed. Nevertheless, as noted in Section 2.2, clues are provided as to what this word refers to: a *veck* is *old* and may be *tolchocked*, a procedure that will result in the *veck* losing blood. What is not revealed, however, by looking at just one instance is the regularity of the phraseology of *veck*. In short, *vecks* are, as here, commonly described as *old* (ten co-occurrences) or as the Nadsat equivalent, *starry* (19). Moreover, as the novel proceeds, *starry* replaces *old*: five of the first seven instances of *veck(s)* co-occur with *old* and seven instances of *old veck* are found in the first two chapters; 12 of the 19 instances of *starry* occur not in Part 1 but in Parts 2 and 3 of *ACO*. This gives an indication of how skilfully Nadsat is introduced into the novel a point not previously noted.

While frequent nouns such as *veck* can give an idea of the workings of Nadsat, exploring the phraseology of verbs can also give an idea of the ways in which the syntax of Nadsat deviates from standard English. This is important because it cannot be revealed by keyword analysis alone. Also, unusual syntax, as seen in the example of DO *the ultra-violent on* in Section 3, is arguably an important aspect of Nadsat (Maher, 2010), but one that is probably not as salient to readers as the Russian lexis in the novel. The verb chosen, *smecking / smeck* (from the Russian noun 'smekh' "(a) laugh"), was first seen in the extract in section 2.

If we just consider *smecking*, the form in the extract, concordance analysis of the 30 instances of this form in the book identified the patterns of usage indicated in Figure 3. Assuming we can substitute *laughing* for *smecking*, some of these uses seem quite normal, for example *smecking at someone* or *smecking one's gulliver* ("head") *off*. Others, however, seem unusual and may be associated with Nadsat syntax. To take one example from Figure 3, it is extremely unusual to use a verb with a meaning like 'laugh' in the pattern GO V-ing followed by an adverbial indicating direction. We could also point out that laughing is not usually followed by 'after someone'.

A final aspect of Nadsat usage and phraseology that is worth considering is how rarer words – those with only one or two occurrences in the text – are contextualised, since for these items there is no option of 'picking up' the meaning through multiple encounters. One such item is *rassoodocks* which appears in the third line of the book and then again, in almost exactly the same wording, in the final chapter. The line in question is 'we sat in the Korova Milkbar making up our rassoodocks what to do with the evening' (Burgess, 2012: 7). In this case, the choice of co-text for *rassoodocks* is very helpful in indicating what this item may mean, principally because the word following MAKE *up* + possessive pronoun is 'mind/minds' in over 80% of cases,[19] a level of association that Stubbs (2006) finds to be highly unusual, and one which helps readers predict what *rassoodocks* refers to with a high degree of probability.

GO *smecking* off / into the room

X (was) *smecking* at us / after me

X (was) *smecking* | all over my rot

my / his gulliver off

fit to crack

like bezoomny

away

smecking young ptitsa / bullies / millicents / malchicks
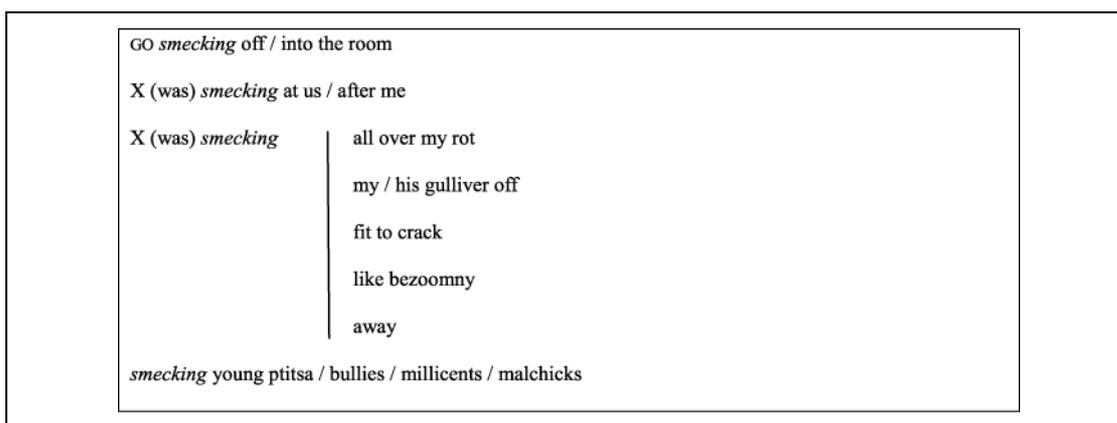
Figure 3. Phraseology of Nadsat *smecking*.

5. Conclusion

As the first rigorous attempt to define and delimit Nadsat, this study has utilised corpus methods in order to identify elements within the text of *ACO* which deviate from standard English. By extracting keywords compared to a very large corpus and carefully investigating candidates, we propose that Nadsat is more extensive than both the in-text definition provided by the character Dr Branom and the

glossaries and commentaries by various scholars over the past five decades. This paper demonstrates that Nadsat can be divided into 'core' items and a series of less frequent categories that commonly replicate adaptations made to Russian transliterations such as reduplication (*malchickiwick*), truncation (*veck*) and wordplay (*horrorshow*). Our analysis suggests that, far from being a mere relexification of Russian into English, Nadsat is a complex creation which functions to render itself comprehensible via a broad range of linguistic and stylistic strategies. We have additionally indicated how techniques of corpus analysis used by other studies of corpus stylistics such as Mahlberg (2013) can help isolate items of interest and unveil the mechanisms at work in reader engagement with invented languages.

Due to limitations of space, it has not been possible to explore in depth the phraseology of Nadsat, but some illustrative analyses have been provided. These offer evidence that a phraseological perspective on Nadsat can not only demonstrate the skill with which items are introduced but also provide insight into other syntactic aspects of Nadsat that are less salient but arguably contribute to its distinctiveness.

A further limitation is that we have not been able to discuss the potential importance of metaphor to Nadsat; coinages such as "the till's guts" (see extract in section 2.2) can be viewed as functions of literary style while still having a relationship with the invented language component of the novel. Indeed, this points to the limitations of keyword analysis. The prevalence of onomatopoeic terms, which relates in part to the importance of Russian-derived sounds noted by some commentators, is likewise an issue worthy of further exploration.

Because Nadsat is not a full art language, it is translatable and indeed has been translated into around 32 languages, but in order to understand how it is possible to translate a non-natural language it was first important to identify what the language is and how it functions. We conclude that Nadsat is an artistically created anti-language, with a core lexis of mostly Russian derivation, augmented by a series of smaller linguistic effects which we have quantified and examined. In future studies we hope to examine the strategies used by translators seeking to migrate Burgess's invention into various target languages, a task already started by Malamatidou (2017).

Acknowledgements

Funding

Notes

1. Also known as 'artlangs', these are languages produced for artistic purposes, for example the Elvish languages in the work of Tolkien or, more recently, the languages invented for the *Game of Thrones* series (Peterson, 2015). Art languages are a sub-type of constructed languages, or conlangs (e.g. Esperanto).
2. In this paper we use the following conventions:
   italics – Nadsat words
   SMALL CAPS – lemmas
   "double speech marks" – glosses
3. As Evans (1971) notes, this claim then appeared on the dust jacket of the 1962 Heinemann edition of the book.
4. Evidence supporting this claim has been provided by Saragi et al. (1978) for adult English-speakers and Pitts et al. (1989) from English-as-a-second-language speakers.
5. Key to Nadsat items in extract:
   deng = "money"
   crasting = "stealing"
   pretty polly = "money" (rhymes with *lolly*, a slang word for money)
   tolchock = "beat up"
   veck = "person, man"
   viddy = "see, watch"
   starry = "old"
   ptitsa = "bird (i.e. woman)"
   smecking = "laughing"
6. There are eight Russian-derived items in the 76-word extract, which equates to 1052 per 10,000 words; in the book overall the figure is 584 per 10,000 words. This may in fact explain why this passage is commonly used (e.g. Maher, 2010) to illustrate the use of Nadsat.

7. e.g. *hen-korm* ("chicken-feed") is classed as non-Russian, when *korm* clearly has a Russian root; *prod* is glossed as "produce" when the Russian cognate means "sell".
8. Biswell has disowned this glossary as it was produced by the publisher of the Restored Edition of *ACO* and not by him. We refer to it in this manner so as to distinguish between the multiple glossaries and not to attribute the compilation to him.
9. In fact, Gabrielatos and Marchi (2011) show over a range of corpora that there is little or no relationship between a keyword list produced using this measure and one produced using their effect size measure, which calculates the size of the difference in terms of percentage difference in frequency between target corpus and reference corpus.
10. Sketch Engine is an online interface including a suite of tools which can be used for corpus analysis. It contains a large number of ready-to-use corpora such as En Ten Ten (see note 11) but also allows users to upload and tag their own corpora for parts of speech; such corpora can then be compared with those already loaded on the interface, making it very useful for our purposes.
11. The En Ten Ten corpus (standing for English 10^{10} i.e. 10 billion words) is a 19-billion word general corpus, collected from a wide range of websites in English and cleaned using automated software (see https://www.sketchengine.co.uk/documentation/tenten-corpora/ for further details (accessed 1 April 2017). This corpus was used as a reference corpus partly due to its size, a factor considered important by Leech and Short (2007: 42). More importantly, this general corpus was chosen as a reference corpus rather than a corpus of fictional works because the reference corpus should be representative (inasmuch as this is possible – Mahlberg, 2013) of the variety that one is comparing against (here, standard English). Using a corpus of literary works would help us identify what is not *literary* about the language of the book, not what is not standard.
12. Other expressions occurring in the book are *as they called them*; *as he/it was called*; *what was called*. Other items excluded in this way include *charlie/charles, polyclef* and *statemart*.
13. The full list of Nadsat words can be found at https://ponyingtheslovos.wordpress.com/2017/04/20/breaking-down-nadsat-into-categories/ (accessed 25 April 2017).
14. The pair *droog, droogie* offers evidence of the advisability of separating these words. Some previous works have assumed that *droogie* is a diminutive form of *droog* but in fact it is also used as an adjective/adverb in the book ('smiling very wide and droogie', Part 1 Chapter 2).
15. There is not enough space to go into detail about how these meanings have been worked out, but in essence this is from looking at the context in which they are used. For example, the meaning of *twenty-to-one* is apparent in its consistent use in contexts such as 'this would sharpen you up and make you ready for a bit of dirty twenty-to-one' (Burgess, 2012: 7) where the rhyming 'fun' could clearly be substituted for *twenty-to-one*.
16. The fact that the rhyming equivalent of this is a taboo word may explain why Burgess did not write the full rhyming slang couplet in this case.
17. It is interesting to note that this figure (just under 6%) closely corresponds to Laufer's (1989) estimate of the proportion of unknown words that a reader of a text can reasonably cope with.
18. We identified the following: *or*; *as we called it*; *that is*; *I mean*; *we called them*; *is what we call*.
19. Based on searches carried out in the British National Corpus.

References

Annaud J-J (1981) *Quest for Fire* [film]. Canada/France.
Biswell A (2012) Introduction. *A Clockwork Orange: The Restored Edition*. London: William Heinemann.
Burgess A (1964) *Language Made Plain*. London: *Hodder & Stoughton Ltd.*
Burgess A (1972) Clockwork marmalade. *The Listener* 87 (2238), 17 February 1972, 197–199.
Burgess A (1990) *You've Had Your Time: Being the Second Part of the Confessions of Anthony Burgess.* London: Heinemann.
Burgess A (1992) *A Mouthful of Air*. London: Hutchison.
Burgess A (2012) *A Clockwork Orange: The Restored Edition (ed.* Andrew Biswell). London: William Heinemann.
Carson, J. (1976) Pronominalization in *A Clockwork Orange*. *Papers on Language and Literature* 12(2): 200-205.
Dunning T (1993) Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1): 61–74.
Evans RO (1971) Nadsat: The argot and its implications in Anthony Burgess' *A Clockwork Orange*. *Journal of Modern Literature* 1(1): 406–410.
Fowler R (1979) Anti-languages in fiction. *Style* 13(3): 259–278.
Gabrielatos C and Marchi A (2011) Keyness: Matching metrics to definitions. In: *Corpus linguistics in the south: Theoretical-methodological challenges in corpus approaches to discourse studies – And some ways of addressing them*, University of Portsmouth, UK, 5 November 2011.
Halliday MAK (1976) Anti-languages. *American Anthropologist* 78(3): 570–584.
Hyman SE (1963) Afterword. *A Clockwork Orange.* New York: Norton.
Jackson H (2011) Invented vocabularies: The cases of Newspeak and Nadsat. In: Adams M (ed.) *From Elvish to Klingon: Exploring Invented Languages.* Oxford: Oxford University Press, pp.49–73.
Jackson K (1991) Real horrorshow: A short lexicon of Nadsat. *Sight and Sound* 9(9): 24–27.

Johnson J (2016) A comparable comparison? A corpus stylistic analysis of the Italian translation of Julian Barnes' *Il Senso di una Fine* and the original text *The Sense of an Ending*. *Language and Literature* 25(1): 38–53.

Kilgarriff A (2009) Simple maths for keywords. In: Mahlberg M, González-Díaz V and Smith C (eds) *Proceedings of Corpus Linguistics 2009*. Available at: http://ucrel.lancs.ac.uk/ publications/cl2009/ (accessed 10 September 2016)

Kilgarriff A (2012) Getting to know your corpus. In: Sojka P, Horak A, Kopecek I and Pala K (eds) *International conference on text, speech and dialogue*, *September 3-7, 2012, Brno, Czech Republic*, pp.3–15. Berlin/Heidelberg: Springer.

Kohn L (2008) Antilanguage and a gentleman's Goloss: Style, register, and entitlement to irony in A Clockwork Orange. *eSharp* 11: 1-27. Available at http://www.gla.ac.uk/media/media_81291_en.pdf (accessed 10 September 2016)

Laufer B (1989) What percentage of text lexis is essential for comprehension? In: Lauren C and Nordman M (eds) *Special Language: From Humans Thinking to Thinking Machines*. Clevedon: Multilingual Matters, pp.316–323.

Leech GN and Short MH (2007) *Style in Fiction: A Linguistic Introduction to English Fictional Prose.* 2nd ed. London: Longman.

Lennon B (2010) *In Babel's Shadow: Multilingual Literatures, Monolingual States*. Minneapolis, MA: University of Minnesota Press.

Maher B (2010) Attitude and intervention: *A Clockwork Orange* and *Arancia meccanica*. *New Voices in Translation Studies* 6: 36–51.

Maher B (2011) *Recreation and Style: Translating Humorous Literature in Italian and English*. Philadelphia, PA: John Benjamins.

Mahlberg, M. (2013) *Corpus stylistics and Dickens's fiction*. Abingdon: Routledge.

Malamatidou S (2017) Creativity in translation through the lens of contact linguistics: A multilingual corpus of A Clockwork Orange. *The Translator* 23(2): 1–18.

Mukařovsky J (1958) Standard language and poetic language. In: Garvin PL (ed. and tr.) *A Prague School Reader on Esthetics, Literary Structure and Style.* Washington DC: American University Language Centre, pp.17-30

Noon G (1993) *Vurt*. London: Pan Macmillan.

O'Halloran K (2007) Corpus assisted literary evaluation. *Corpora* 2(1): 33–63.

Partridge E (1937) *A Dictionary of Slang and Unconventional English.* London: Routledge.

Partridge E and Beale P (1984) *A Dictionary of Slang and Unconventional English* (8th edn.)*.* London: Routledge.

Peterson, D. (2015) *The Art of Language Invention*. New York: Penguin.

Pitts M, White H and Krashen S (1989) Acquiring second language vocabulary through reading: A replication of the Clockwork Orange study using second language acquirers. *Reading in a Foreign Language* 5(2): 271–275.

Rawlinson M (2011) *A Clockwork Orange: The Norton Critical Edition*, ed. Mark Rawlinson. New York: Norton.

Rayson P (2013) Corpus analysis of key words. In: Chappelle C (ed.) *The Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell.

Saragi T, Nation ISP and Meister GF (1978) Vocabulary learning and reading. *System* 6(2): 72–78.

Scott M (1997) PC analysis of key words – And key key words. *System* 25(2): 233–245.

Sinclair J (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stubbs M (2005) Conrad in the computer: Examples of quantitative stylistics methods. *Language and Literature* 14(1): 5–24.

Stubbs M (2006) Corpus analysis: The state of the art and three types of unanswered questions. In: Thompson G and Hunston S (eds) *System and Corpus: Exploring Connections*. London: Equinox, pp.15–36.