*Henk Barkema**

The Idiomatic, Syntactic and Collocational Characteristics of Received NPs: some basic statistics

Introduction[1]

Despite the growing interest that researchers from various linguistic fields have had in the syntax of idioms, proverbs, binomials and other types of 'received expression' during the last decade, no one has yet been able to provide an exhaustive description of grammatical 'flexibility': the degree to which (and the way in which) the forms of received expressions can be varied morphologically and syntactically.[2] One thing is evident: the flexibility of received expressions is limited compared with that of free expressions; the question is, however, to what extent.

Our research objective is firstly to give a thorough and systematic description of the difference in flexibility between received and free expressions and, secondly, to describe what the relation is between the limited flexibility of received expressions and other types of characteristics of these expressions.

In order to reach this aim we are planning a large-scale statistical comparison of the flexibility characteristics of received expressions with those of free expressions. In order to be able to make such a comparison we collected the occurrences of a large number of received expressions from a relatively large balanced corpus: the 20 million Birmingham Collection of English Texts (see Renouf, 1984). For prac-

[1] I am most grateful to Jan Aarts for his comment on earlier versions of this article and to Inge de Mönnink for her help with the figures.

[2] Received expressions are expressions with one or more idiosyncratic characteristics that have to be learnt by language learners. As lexicalized expressions they are given special attention in dictionaries.

* *Henk Barkema*
TOSCA Research Group
Department of Language and Speech
University of Nijmegen
The Netherlands

tical reasons we restricted ourselves to the examination of received noun phrases, although the methodology that we apply to this type of phrase can be easily adapted to other types as well as to clauses and sentences. Besides, we take it that some of the results found in relation to NPs can be used for the formulation of hypotheses about the flexibility characteristics of other types of received expression.

The set of received NPs that we are examining consists of all NPs that are listed in the Longman Dictionary of Contemporary English (Procter, 1978), compounds and ‘loan expressions’ (such as *joie de vivre*) excluded. The total number of this set is 2,185.

We collected all occurrences of these expressions from the Birmingham Collection of English Texts (henceforth: Birmingham corpus), after which every occurrence was annotated with detailed information about its morpho-syntactic form. This procedure is described in Barkema (1993). In a later stage we provided all occurrences with codes indicating the compositionality, collocability, formulaicity as well as the syntactic structures of the base forms of the expressions. In addition the function of the occurrence in the sentence, the presence of an initial determiner and coordination with other phrases was indicated. The result was for each occurrence a string of codes that was translated into a line of ones and zeros where each code had its own position. The resulting matrix file contained a total of 50,670 lines of about 500 columns.

In this paper we first give a brief description of the terms that we use to refer to the various characteristics of received expressions and then report on some first results that were derived from our annotated collection.

The Forms of Received Expressions

The base form of a received expression is the simplest form that it takes. It is therefore the form which can be found in the dictionary. If the base form of the received expression contains one or more empty function slots we call it a template form. Examples are *CONJOIN and all that jazz* and *take the law in ONE’S own hands*. If on the other hand all the function slots of the base form of a received expression are filled we call it a minimal form. Examples are *rolled gold* and *pins and needles*. In other words: there are two types of base forms: template forms and minimal forms.

To be able to refer to a single morphological or syntactic alteration of the base form we use the term ‘variation’. A form containing one or more variations we call ‘variant form’. Thus the following variant (in italics) contains only one variation (underscored)[3]:

1) If you ask *the man in the modern street* for his opinion, he would probably ...

The following variant form contains two variations (underscored)[4]:

2) *a hard-fought foot in the door of one of the aviation world’s least exploited markets.*

We call the syntactic structures of the base forms of received expressions ‘base patterns’. Various different expressions have the same base pattern. For example *battle royal*, *mother superior*, *notary public*, *heir presumptive* and *concert grand* all have singular nouns functioning as heads and absolute adjectives functioning as postmodifiers.

Compositionality

We distinguish between four types of compositionality: fully compositional, pseudo-compositional, partly compositional and non-compositional.

Fully compositional expressions have meanings that are completely inferred from the combination of the basic, derived or extended senses of their lexical items and their syntactic structures. Examples are the adjuncts *to put it bluntly* and *to be more precise*.

Pseudo-compositional expressions have meanings that are only partly inferred from the basic senses of their lexical items and their syntactic structures: the other parts are not accounted for. Examples are:[5]

3) *bed and breakfast:* (a private house or small hotel that provides) a place to sleep for the night and breakfast the next morning for a fixed price;

4) *clutch bag:* a type of handbag that is carried in the hand rather than with a strap. It is usu. used by women when they are going out somewhere special in the evening.

[3] This example is from the Birmingham corpus.

[4] This example is from The Economist.

[5] The definitions in this article are from the Longman Dictionary of English Language and Culture (Summers, 1992), unless other references are given.

Non-compositional expressions have meanings that are in no way related to their syntactic structures in combination with the basic senses of their lexical items. They are often referred to as ‘idioms’ or ‘idiomatic expressions’. Examples are:

5) *a finger in every pie:* a part or interest in everything that is going on;

6) *a bitter pill to swallow:* something very unpleasant that one has to accept.

In partly-compositional expressions only parts of their meanings are inferred from only some of the senses of their lexical items. An example is:

7) *broad hint:* a full and clear hint.

Polysemy

If a received expressions has more than one meaning, for example one which is pseudo-compositional and another which is non-compositional, it is polysemous. An example is *bed and breakfast*, which has a second non-compositional meaning:[6]

8) *bed and breakfast:* an operation in which a shareholder sells a holding one evening while agreeing to buy it back again next morning realizing either a gain or a loss in order to suit a tax requirement.

If a received expression is not polysemous, it is monosemous.

Empty senses

A small minority of received NP expressions have lexical items that have no (basic or derived) senses in isolation, in other words: they have lexical items that cannot be found as separate entries with their own meanings in the dictionary; they can only occur as parts of received expressions. Examples are *short shrift* and *high jinks*.

Collocability

A received expression is ‘collocationally closed’ if none of the lexical items from open classes in the expression can be substituted by an alter-

[6] This definition is from Hawkins and Allen's OEED (1991).

native; for example in the received expression *not so dusty* ('fairly good', 'quite well') none of the lexical items can be replaced. Expressions in which a limited number of lexical items can be substituted are 'collocationally limited'. Examples are *a head/memory like a sieve* and *a drop in the bucket/ocean*. If alternatives are possible they are often synonyms or near-synonyms.

Finally, in a number of received expressions a set of semantically related lexical items from an open class can be substituted. Examples are *TIME after TIME*, where at the position of *TIME* a series of near-synonyms can be placed: *day after day, month after month*, *second after second*, etc. We call such expressions 'collocationally open'.

Formulaicity

Some expressions have institutionalized pragmatic functions. If this is the case such special pragmatic information is provided in the dictionary. An example is:

9) *good God/grief/gracious/heavens/Lord!* ('used as an expression of surprise or other strong feeling').

We call such expressions 'formulaic'.

For a more detailed description of the various characteristics of received expressions: see Barkema (in preparation).

The Coding of the Occurrences

In order to be able to discover the nature of the limited flexibility of received expressions each occurrence was provided with a code that indicated of which received NP it was an occurrence. In addition each occurrence was given a location code. Also, we annotated each occurrence in the corpus with information about the various types of characteristics of the expression itself: its compositionality, its collocability and its formulaicity. If lexical items with empty senses were present in the expression or if it was polysemous, this was coded too. Finally, the base pattern of the expression was expressed in the form of a code.

In addition each occurrence was annotated with information about the (base or variant) form of the occurrence. If it was a variant form, detailed information was added about the way in which the variant form differed from the base form by describing each of its variations. Of each variation it was indicated:

a) at which node in the structure of the NP it was found;
b) with which syntactic function it was found;
c) which *type* of variation was found (addition, term selection, permutation or insertion);
d) which word classes were involved in the variation.

For example, if in an occurrence an additional adjective was found as premodifier of the head in a postmodifying PP, this could be read from the code string.

The codes expressing the base patterns of the received NPs at the moment only indicate the syntactic functions of the immediate constituents of the noun phrase and phenomena like apposition and coordination. Although at this stage we have not included information about the realization of the functions, we have made an exception for the realization of the postmodifier: a distinction has been made between clauses and PPs realizing this function.

Finally, in order to be able to examine the relation between the initial determiner and the flexibility of the received NP, the presence of initial determiners and their realizations were coded separately.

A First Impression of the Statistical Data

In the remainder of this article we discuss the frequencies of types and tokens of received expressions in relation to their various types of characteristics, to types of variation, to syntactic functions realized by them, to coordination of received NPs with other phrases and to the realizations of their initial determiners.

With the term ‘token’ we mean the actual occurrences of received expressions in the corpus and with the term ‘type’ the expression abstracted from its occurrences; for example, of the type *dark horse* nine tokens were found in the Birmingham corpus.

1) The Various types of compositionality

Although the best-known received NP expressions are the non-compositional or ‘idiomatic’ ones, only 16.89 per cent of the total number of received NP expressions turned out to be idiomatic (369 out of 2,185 types). The number of tokens of these expressions found in the Birmingham corpus was only 6.37 per cent of the total number of tokens found (3,234 out of 50,760). The group of partly-compositional expressions is slightly bigger: 21.51 per cent of the total number of types (470 out of 2,185), with 10.67 per cent of the total number of tokens (5,419 out of 50,760), while the group of fully compositional received NPs is relatively small: only eight types had 1.25 per cent of the total number of tokens (632 out of 50,760). The majority of received NPs are pseudo-compositional: 61.24 per cent of the total number of types (1,338 out of 2,185) and 81.71 per cent of the total number of tokens (41,475 out of 50,760) were pseudo-compositional - see figure 1.[7]

Figure 1 a

Figure 1 b

Figure 1: the proportions of fully compositional, pseudo-compositional, partly-compositional and non-compositional received NP expressions in the Birmingham corpus.

[7] The order of the items in the pies is clockwise from the top.

It seems that the more compositional received expressions are, the more tokens per type they have on average: fully compositional expressions have the highest number (632/8=79.00 tokens per type), followed by pseudo-compositional (41,475/1,338=31.00 tokens per type), partly compositional (5,419/470=11.53 tokens per type) and non-compositional expressions (3,234/369=8.76 tokens per type). In other words: the more compositional an expression is, the more often it is used.

2) Polysemous versus Monosemous expressions

The majority of expressions turned out to be monosemous: 1,888 out of 2,185 types (86.41%) and 44,954 out of 50,760 tokens (88.56%). There is no big difference in the average number of tokens per type (polysemous: 5,806/297=19.55 and monosemous: 44,954/1,888=23.81 tokens per type); see figure 2.

Figure 2 a

Figure 2 b

Figure 2: the proportions of monosemous and polysemous expressions.

3) Empty sense expressions

The majority of received NPs appeared to have no lexical items with empty senses: 2,156 out of 2,185 types (98.67%) and 50,606 out of 50,760 tokens (99.70%). In addition empty sense expressions seem to have fewer tokens per type than expressions without empty senses (empty sense expressions: 154/29=5.31; no empty sense expressions: 50,606/2,156=23.47); see figure 3.

Figure 3 a

Figure 3 b

Figure 3: the proportions of empty sense expressions and ones that do not have lexical items with empty senses.

4) Collocability

Out of 2,185 expressions 1,592 (72.86%) turned out to be collocationally closed (31,487 tokens, 62.03%), 588 (26.91%) were collocationally limited (18,966 tokens, 37.36%), while 5 (0.23%) were open (307 tokens, 0.61%); see figure 4.

Figure 4 a

Figure 4 b

Figure 4: the proportions of collocationally closed, limited and open expressions.

It seems that the more collocationally open a received expression is, the more frequently it is used: collocationally open received NP expressions had 307/5=61.40 tokens per type on average, collocationally limited ones 18,966/588=32.26 and collocationally closed ones 31, 487/1,592=19.78.

5) Formulaicity

Out of 2,185 expressions 40 (1.83%) with a total of 932 tokens (1.84%) were formulaic, while 2145 (98.17%) were non-formulaic (49,828 tokens, 98.16%); see figure 5.

Figure 5 a

Figure 5 b

Figure 5: the proportions of formulaic and non-formulaic expressions.

On average formulaic expressions occurred as frequently as non-formulaic ones (formulaic expressions: 932/40=23.30 and non-formulaic ones: 49,828/2,145=23.23).

6) Syntactic structures

Although we found a total of 20 different base patterns in the set of NPs, 98.06 per cent of the tokens had one of the base patterns listed in figure 6.

Base Pattern:	Number of types: (total: 2,185)		Number of tokens: (total: 50,760)		Av. no. of tokens per type:
premodifier + head	1,508	(69.01%)	32,208	(63.45%)	21.35
determiner + head	181	(8.28%)	110,250	(20.19%)	56.63
head + postmod. PP	270	(12.36%)	4,379	(8.63%)	16.22
coordination	101	(4.62%)	1,682	(3.31%)	16.65
head + adverbial	41	(1.88%)	775	(1.53%)	18.90
head + postmod. clause	6	(0.27%)	486	(0.96%)	81.00
	------------		-----------		
Total	2107	(96.43%)	49,780	(98.06%)	

Figure 6 a

Figure 6 b

Figure 6: the most frequent base patterns with their frequencies

The other 14 base patterns cover less than 2% of the total number of types and less than 4% of the total number of tokens.

Although they are very infrequent, expressions with the base pattern ‘head plus postmodifying clause’ have quite a large number of tokens per type on average, namely 81.00; the same goes for expressions with the base pattern ‘determiner plus head’ (56.63 tokens per type on average).[8]

[8] The majority of these determiners are possessive ones.

Frequencies of Variations

We distinguish between four types of variation: addition, term selection, permutation and insertion (compare Barkema 1993, 1994). A variation is an ‘addition’ if in the base form of a received expression an additional function occurs. An example is *a constant bone of contention*. A variation is a ‘term selection’ if from a closed class an alternative has been selected, for example: singular/plural variation, absolute/comparative/superlative variation or definite/indefinite article variation. Examples are *sitting ducks* and *the straw that broke the camel’s back*. A variation is a ‘permutation’ if the order of the elements in the variant form is different from that in the base form, e.g.: *heir presumptive* versus *presumptive heir.*[9] Finally, a variation is an ‘insertion’ if an additional function interrupts the syntactic structure of the base form of the expression. An example is *the straw that -finally- breaks the camel’s back.*

If two similar variations were found within the same function and at exactly the same position in the structure of the phrase, we counted them as one, for example two additional adjectives premodifying the head of the NP.

The Various Classes of Variation and their Frequencies

Of the total of 50,760 tokens 31,621 (62.3%) turned out to be base forms. The other 19,139 tokens (37.7%) contained one or more (classes of) variations.

In 13,128 tokens (25.86% of the total number of tokens) additions were found, in 7,497 tokens (14.77%) term selections, in 724 tokens (1.43%) permutations and in 63 tokens (0.12%) insertions.[10]

In the 19,139 tokens with variations we found 23,530 variations: 15,020 additions (63.83% of the total number of variations), 7,717 term selections (32.80%), 729 permutations (3.10%) and 64 insertions (0.27%); see figure 7.

[9] It is not important which of these forms is regarded as the base form as long as it is indicated when permutation has been found.

[10] The total number of these tokens is 21,412 (and not 19,139). The reason for this is that every time a token contains two classes of variations, it is counted twice.

Figure 7 a

Figure 7 b

Figure 7: the distribution of the variation types of variations over the variant forms.

In table 1 we show how many different variations were found per variant form. Only addition and term selection appear to have more than two variations per variant form in the corpus.

	1 variation:	2 variations:	3 var.: >	3 var.:	total:
addition	11,369 (75.69%)	1,635 (10.89%)	115 (0.77%)	9 (0.06%)	15,020
term sel.	7,292 (94.49%)	192 (2.49%)	112 (0.16%)	1 (0.01%)	7,717
permutation	719 (98.63%)	5 (0.69%)	-	-	729
insertion	62 (96.88%)	2 (3.12%)	-	-	64

Tabel 1: the numbers of different variations per variant form.

Syntactic Functions of the Expressions

In figure 8 we list the functions realized by the received NP expressions, with their frequencies.

type of syntactic function:	number of tokens: (total: 50,760)
prepositional complement	20,126 (39.6%)
adverbial	6,595 (13.0%)
subject	6,183 (12.2%)
direct object	5,879 (11.6%)
premodifier	4,179 (8.2%)
subject complement	2,625 (5.2%)
no function (utterance)	1,583 (3.1%)
apposition	909 (1.8%)
head of possessive determiner phrase	365 (0.7%)
determiner	337 (0.7%)
vocative	240 (0.5%)
object complement	187 (0.4%)
postmodifier	185 (0.4%)
reduced clause	163 (0.3%)
verb modifier	65 (0.1%)

NB: an example of the function 'head of possessive determiner phrase' is: *the United Nations' program for the future*.

other types of function: < 0.1%

Figure 8

Figure 8: the functions of received NP expressions.

The functions most often realized by received NPs are prepositional complement (39.6% of the tokens), adverbial (13.0%), subject (12.2%) and direct object (11.6%). It is interesting to see that the function of adverbial is realized by received NPs more frequently than those of subject and direct object.

Coordination of Received Expressions

In figure 9 we can see that received NPs are hardly ever coordinated with other phrases. Only 10.0% of the tokens were found to be coordinated (5,076 out of 50,760). If received NPs were coordinated, this was most often found to be with NPs (4,163 tokens, 82.0%) and less frequently with AjPs (61 tokens, 1.2%). In addition, received NP expressions turned out to be coordinated more frequently with free NPs (3,068 tokens, 60.4%) than with received NPs (749 tokens, 14.8%). A number of received NPs that were found to be coordinated with other phrases shared the determiners or postmodifiers with these phrases.

Figure 9

specification of the coordination with NPs:

with a free NP	3,068 (60.4%)
with a received NP	749 (14.8%)
with a free NP (shared determiner)	261 (5.1%)
with a rec. NP (shared determiner)	56 (1.1%)
with a free NP (shared postmodifier)	29 (0.6%)

Figure 9: coordination of received NPs.

Additional Initial Determiners of Received Expressions

As we said in the first part of this article, the presence and realization of additional initial determiners was expressed in the form of a separate code. If the initial determiner function of the base form of a received NP was realized by a category other than that of (definite or indefinite) article, this category was regarded as part of the base form, e.g. *dead man's handle*, *all the vogue*. Categories that were found to realize the function of initial determiner and that did not belong to the base forms of received expressions were regarded as additional ones, e.g. *a clean slate*, *this full stop*.

In figure 10 we can see that most of the tokens had no additional initial determiners (41.1%). It is interesting to see that in 31.9% of the cases no initial determiner was found, although the noun functioning as the head of the NP was a singular noun. However, we have to be careful here: we have not yet distinguished between mass and count nouns. Nevertheless we came across many singular count nouns without an additional determiner and we have the impression that the combination of singular nouns with zero determiners is typical of received NPs.

The majority of the additional initial articles that we found appeared to be definite ones 30.7%), while relatively few were indefinite (8.2%).

Type of determiner:	Number of tokens: (total: 50,760)
absent determiner	20,864 (41.1%)
article	19,745 (38.9%)
initial determiner in base form	5,594 (11.0%)
added pronoun	3,422 (6.8%)
quantifying group	497 (1.0%)
possessive group	478 (0.9%)

other types of determiner: < 1%

Figure 10

specification of absent determiners:

absent (noun-head is singular count noun or mass noun)	16,217 (31.9%)
absent (noun-head is plural)	4,647 (9.2%)

specification of types of additional articles:

definite article	15,603 (30.7%)
indefinite article	4,142 (8.2%)

specification of types of additional pronouns:

possessive pronoun	2,268 (4.5%)
quantitative pronoun	272 (0.5%)
demonstrative pronoun	291 (0.6%)
negative pronoun	221 (0.4%)
assertive pronoun	192 (0.4%)
non-assertive pronoun	89 (0.2%)
universal pronoun	63 (0.1%)
interrogative pronoun	26 (0.1%)

Figure 10: the types of initial determiner taken by received NPs.

Conclusion

In this article we have reported on some first results derived from an annotated collection of 2,185 received NPs with a total of 50,670 tokens. The information we have now available will be used to compare the characteristics of the received NPs with those of free NP expressions. In Barkema (1994) we have described a method by means of which an inventory of NP structures from a parsed corpus can be used to statistically compare the flexibility characteristics of received expressions with those of free expressions.

References

Barkema, H. (1993): 'Idiomaticity in English NPs'. In: J. Aarts, P. de Haan and N. Oostdijk: *English Language Corpora: Design, Analysis and Exploitation. Papers from the thirteenth international conference on English language research on computerized corpora, Nijmegen, 1992.* Amsterdam: Rodopi.

Barkema, H. (1994): 'Determining the Syntactic Flexibility of English Idioms'. In: G. Tottie and U. Fries: *English Corpus Linguistics. Papers from the 14th ICAME conference in Zürich.* Amsterdam: Rodopi.

Barkema, H. (in preparation): *Idiomaticity and Terminology: a Multi-Dimensional Descriptive Model.*

Hawkins, J. and R. Allen (1991): *The Oxford Encyclopedic English Dictionary*. Oxford: Clarendon Press.

Procter, P. (1978): *The Longman Dictionary of Contemporary English*. Harlow: Longman Group Ltd.

Renouf, A. (1984): 'Corpus development at Birmingham University', in J. Aarts and W. Meijs: *Corpus Linguistics II: New Studies in the Analysis and Exploitation of Computer Corpora.* Amsterdam: Rodopi.

Summers, D. (1992): *Longman Dictionary of English Language and Culture*. Harlow: Longman Group Ltd.