The Development of a Functional Function Word List from *Form* and *Meaning*

Warren TANG*

かたち
文字と意味から応用のための機能語リスト

タン　ウォーレン*

ABSTRACT

The paper proposes a novel way to produce an English function word list by using not only form but also meaning by way of parts-of-speech. Rather than simply having a straight list of words, words are divided into *function/content form* and *function/content meaning*. The results show that 43.4% of function words have function form and function meaning, 2.2% have function form and content meaning, 1% have content form and function meaning, and 53.4% have content form and content meaning. The list of 387 word-forms developed out of this data showed that it covers 43.3% of the possible 44.4%, just 1.1% short of full coverage of function words, and is 2.5% better than currently available, comparable lists.

Keywords: function word, corpus, wordlist, form, meaning, parts-of-speech

1. Introduction

Grammarians have given function words a bad name. Charles Fries (1977), who gave us this term, had noted that function words began life with a different name – *empty words*[i]. While today we have retained the term for its other, "better" half – *content words*[ii] – we are always still left with a feeling that function words are *words without content*.

However, to define function words it is still best to first define what content words are. Consider the previous sentence. If we retain the nouns, verbs, adjectives, and adverbs (*define*, *function*, *words*, *still*, *best*, *first*, *define*, *content*, *words*) we can still guess at its meaning. Retain the opposite set of words (*so*, *to*, *it*, *is*, *to*, *what*, *are*) and you are left with something most people would consider gibberish. In a sense naming nouns, verbs, adjectives, and adverbs content words is justified. But a text or sentence with only content words or function words would not make for a readable text. Both sets of words are necessary. To paraphrase Biber's metaphor, lexical words (content words) are the building blocks (bricks) of text and function words are the mortar which hold the text together (Biber et al., 1999). In other words, a text with content words alone would not be a very good piece of communication. A text is a combination of content and function words. And a good text is one which combines content and function words well. It is best to define function words negative terms (negative not in meaning, but by non-affirmative means). The definition used here for function words is words which are not nouns, verbs, adjectives, or adverbs.

Having said this, then, we must assess what is left. Traditionally, prepositions, pronouns, determiners, and conjunctions fell under the category of function words. But this leaves many other parts-of-speech (POS) in limbo. With our simple definition then all else – *modal auxiliary verbs*, *primary auxiliary verbs*[iii], *(adverb) particles*, *question words*, *number words*, *the possessive marker 's'*, and *the infinitive marker 'to'*[iv] – will be all considered function words, with the exceptions of *punctuation*, *alphabetic letters*, and *unclear instances* to be classified as content words.

*大学教育センター助教

Our definition of a word is uppercase [A-Z] and lowercase [a-z] letters of the alphabet unless stated otherwise.[v]

1.1 Why focus on function words?

Even though, as demonstrated above, language cannot function without function words, the general characterisation of function words as to meaningless goes against common sense. The move towards a *lexicogrammar* paradigm where morphology and syntax[vi] are not seen as two separate components but as inseparable and necessary[vii]. Thus, focusing on function words really is not pitting words groups or classes against each other but rather highlighting that they are not so different. While they differ in characteristics, they are not different in being a part of language.

2. Meaning

For the construction of the list we have used the data from the BNCweb (2012).[viii] Below is a list of all the POS tags (CLAWS 5 tagset), their definitions, and the additional tag used in this study[ix] (F for *function meaning* and C for *content meaning*).

Table 1 – POS tags of BNC/BNCweb (CLAWS 5 tagset) and supplemental tag

POS	*Tag*	*Definition*
AJ0	C	adjective (unmarked) (e.g. GOOD, OLD)
AJC	C	comparative adjective (e.g. BETTER, OLDER)
AJS	C	superlative adjective (e.g. BEST, OLDEST)
AT0	F	article (e.g. THE, A, AN)
AV0	C	adverb (unmarked) (e.g. OFTEN, WELL, LONGER, FURTHEST)
AVP	F	adverb particle (e.g. UP, OFF, OUT)
AVQ	F	wh-adverb (e.g. WHEN, HOW, WHY)
CJC	F	coordinating conjunction (e.g. AND, OR)
CJS	F	subordinating conjunction (e.g. ALTHOUGH, WHEN)
CJT	F	the conjunction THAT
CRD	F	cardinal numeral (e.g. 3, FIFTY-FIVE, 6609) (excl ONE)
DPS	F	possessive determiner form (e.g. YOUR, THEIR)
DT0	F	general determiner (e.g. THESE, SOME)
DTQ	F	wh-determiner (e.g. WHOSE, WHICH)
EX0	F	existential THERE
ITJ	F	interjection or other isolate (e.g. OH, YES, MHM)
NN0	C	noun (neutral for number) (e.g. AIRCRAFT, DATA)
NN1	C	singular noun (e.g. PENCIL, GOOSE)
NN2	C	plural noun (e.g. PENCILS, GEESE)
NP0	C	proper noun (e.g. LONDON, MICHAEL, MARS)
ORD	F	ordinal (e.g. SIXTH, 77TH, LAST)
PNI	F	indefinite pronoun (e.g. NONE, EVERYTHING)
PNP	F	personal pronoun (e.g. YOU, THEM, OURS)
PNQ	F	wh-pronoun (e.g. WHO, WHOEVER)
PNX	F	reflexive pronoun (e.g. ITSELF, OURSELVES)
POS	F	the possessive (or genitive morpheme) ‘S or ‘
PRF	F	the preposition OF
PRP	F	preposition (except for OF) (e.g. FOR, ABOVE, TO)
PUL	C	punctuation – left bracket (i.e. (or [ )
PUN	C	punctuation – general mark (i.e. . ! , : ; – ? …)

PUQ	C	punctuation – quotation mark (i.e. ` ‘ ”)
PUR	C	punctuation – right bracket (i.e.) or])
TO0	F	infinitive marker TO
UNC	C	“unclassified” items which are not words of the English lexicon
VBB	F	the “base forms” of the verb “BE” (except the infinitive), i.e. AM, ARE
VBD	F	past form of the verb “BE”, i.e. WAS, WERE
VBG	F	-ing form of the verb “BE”, i.e. BEING
VBI	F	infinitive of the verb “BE”
VBN	F	past participle of the verb “BE”, i.e. BEEN
VBZ	F	-s form of the verb “BE”, i.e. IS, ‘S
VDB	F	base form of the verb “DO” (except the infinitive), i.e.
VDD	F	past form of the verb “DO”, i.e. DID
VDG	F	-ing form of the verb “DO”, i.e. DOING
VDI	F	infinitive of the verb “DO”
VDN	F	past participle of the verb “DO”, i.e. DONE
VDZ	F	-s form of the verb “DO”, i.e. DOES
VHB	F	base form of the verb “HAVE” (except the infinitive), i.e. HAVE
VHD	F	past tense form of the verb “HAVE”, i.e. HAD, ‘D
VHG	F	-ing form of the verb “HAVE”, i.e. HAVING
VHI	F	infinitive of the verb “HAVE”
VHN	F	past participle of the verb “HAVE”, i.e. HAD
VHZ	F	-s form of the verb “HAVE”, i.e. HAS, ‘S
VM0	F	modal auxiliary verb (e.g. CAN, COULD, WILL, ‘LL)
VVB	C	base form of lexical verb (except the infinitive) (e.g. TAKE, LIVE)
VVD	C	past tense form of lexical verb (e.g. TOOK, LIVED)
VVG	C	-ing form of lexical verb (e.g. TAKING, LIVING)
VVI	C	infinitive of lexical verb
VVN	C	past participle form of lex. verb (e.g. TAKEN, LIVED)
VVZ	C	-s form of lexical verb (e.g. TAKES, LIVES)
XX0	F	the negative NOT or N’T
ZZ0	C	alphabetical symbol (e.g. A, B, c, d)

There are several categorical problems that need to be considered. First, cardinal numbers and ordinal numbers compared to other function word POSs are relatively open class. Particularly, the problem stems from differences between spoken and written forms especially when we are working within a corpus linguistic paradigm. For example, in written form the numeral ‘one’ can be represented by the spelt-out word form or as the numeral ‘1’. Spoken there is no difference in phonological sound form. Furthermore, spelling out a word means we can limit the number of forms necessary to represent all the numerals. Only sixty-eight word forms are needed to represent all the cardinal and ordinal numbers. But numbers without spaces is a single unit, and therefore open to any combination. In other words, one-million numbers is represented by one-million forms in numbers. But with words ‘one million’ needs only the two words of ‘one’ and ‘million’.

Second, a verb with an adverb particle (for example, ‘to sit up’) as a single unit representing a verb. While most of the particles are prepositions their larger role is to aid the main verb in creating a verb more specific meaning (‘sit’ and sit up’ are different verbs with different meanings). Because the forms are strongly associated with the prepositions they will be treated as such.

Third, punctuations are neither function words nor content words. Again, they do not have spoken equivalence. Here they will be ignored as function words and placed into the content words category.

Fourth, ‘unclassified’ and ‘error’ cases will be considered content words. The likelihood that instances are content words is greater than them being function words.[x]

Fifth, primary auxiliary verb forms even though they may also play the role as main verbs. Its content word aspect is ignored, and all instances are classified as function words.

2.2 "Problems" with the corpus linguistic approach

Corpus linguistics is largely a written form-based approach delineated by the spaces and punctuations within texts. Negative forms such as 'doesn't' and 'wouldn't' are counted two separate types – 'doesn' and 'wouldn', and 't'. Within a machine-parsed paradigm these are treated a 'does' and 'would', and 'not'. Either methods are acceptable since the count does not change.

3. Form

From the resulting lists selections were made on based upon mainly frequency (greater than 1%) but also how a type's relationship with other similar types. For example, the plural reflexive 'yourselves' was included because of its close relationship with 'yourself' even though only the latter had a high frequency. Hyphenated words (for example: *vice-president*) were excluded on the letter-only definition even though the BNCweb counts these as one word (type). The result was a list of 387 words (types).

a, aah, aboard, about, above, according, across, after, against, aged, ah, aha, alas, albeit, all, along, alongside, although, am, amen, amid, amidst, among, amongst, an, and, another, any, anybody, anyone, anything, are, aren, around, as, at, aye, back, be, because, been, before, behind, being, below, beneath, beside, besides, between, beyond, billion, billionth, blah, both, but, by, bye, can, cannot, cheers, concerning, considering, cos, could, couldn, crap, d, damn, dare, dear, despite, did, didn, do, does, doesn, doing, don, done, down, during, each, eh, eight, eighteen, eighteenth, eighth, eightieth, eighty, either, eleven, eleventh, enough, every, everybody, everyone, everything, except, excluding, farewell, few, fewer, fewest, fifteen, fifteenth, fifth, fiftieth, fifty, first, five, following, for, former, fortieth, forty, four, fourteen, fourteenth, fourth, from, goddamn, goodbye, goodnight, gosh, ha, had, hadn, half, has, hasn, have, haven, having, he, hello, her, hers, herself, hey, hi, him, himself, his, hiya, hmm, ho, how, however, huh, hundred, hundredth, hurray, hush, i, if, immediately, in, including, inside, into, is, isn, it, its, itself, last, latter, least, less, lest, like, little, ll, lots, m, many, may, me, mhm, might, mightn, million, millionth, mine, minus, more, most, much, must, mustn, my, myself, nah, nay, nd, near, need, needn, neither, next, nine, nineteen, nineteenth, ninetieth, ninety, ninth, no, nobody, none, nope, nor, not, nothing, notwithstanding, of, off, oh, ok, okay, on, once, one, oneself, onto, ooh, oops, opposite, or, ouch, ought, oughtn, our, ours, ourselves, out, outside, over, own, past, pending, per, plenty, plus, provided, providing, rd, re, regarding, right, round, s, same, second, seven, seventeen, seventeenth, seventh, seventieth, seventy, several, shall, shalt, shan, she, should, shouldn, since, six, sixteen, sixteenth, sixth, sixtieth, sixty, so, some, somebody, someone, something, st, such, supposing, t, ten, tenth, th, than, that, the, thee, their, theirs, them, themselves, there, these, they, third, thirteen, thirteenth, thirtieth, thirty, this, those, thou, though, thousand, thousandth, three, through, throughout, thru, thy, til, till, to, toward, towards, trillion, trillionth, twelfth, twelve, twentieth, twenty, two, uh, um, under, underneath, unless, unlike, until, unto, up, upon, urgh, us, used, ve, versus, via, vice, vs, was, wasn, we, well, were, weren, what, whatever, whatsoever, when, whenever, where, whereas, whereupon, wherever, whether, which, whichever, while, whilst, who, whoever, whom, whose, why, will, with, within, without, won, worth, would, wouldn, wow, ye, yeah, yep, yes, you, your, yours, yourself, yourselves, yum, zero

Words like 'don't' are treated as 'don' and 't' as expected in a non-machine-parsed text. However, BNCweb uses machine-parsed data. Therefore, the list was expanded by a further 31 words to include forms with apostrophes. 'don't' is therefore 'do' and 'n't'. And '~'d' was added for analysis purposes. Also 'n' was included to capture 'dunno' due to again quirky machine parsing. In a concordancer like AntConc these words would be counted as 'don' and 't'. But whichever way we count them we are still counting them as two words. Consistency is impossible between corpus systems therefore it is better to cover all forms than to ignore them.

'd, 'll, 'm, 're, 's, 've, aren't, ca, can't, couldn't, didn't, doesn't, don't, hadn't, hasn't, haven't, isn't, it's, mightn't, mustn't, n, n't, needn't, oughtn't, shan't, shouldn't, wasn't, weren't, wo, won't, wouldn't

4. Meaning and form

Combining meaning and form data we can have four possible groups: *function form with function meaning* (FF); *function form with content meaning* (FC); *content form with function meaning* (CF); and *content form with content meaning* (CC). This can be conceptualized in the following manner.

	Function meaning	*Content meaning*
Function form	**FF**	**FC**
Content form	**CF**	**CC**

With this design, what we see is this:

BNCweb (tokens)	*Function meaning (-F)*	*Content meaning (-C)*
Function form (F-)	43.4%	2.2%
Content form (C-)	1.0%	53.4%

For 96.4 percent of the words (tokens), form and meaning match (FF+CC), and 3.2 percent do not (CF+FC). While we have assumed that function and content words are distinct groups there is a small amount of overlap. Of this overlap, FC (2.2 percent) has a larger amount to CF (1.0 percent) comparatively.

We can also say that 44.4 percent of words (tokens) have function meaning (FF+CF) and 55.6 percent have content meaning (FC+CC). This is a good estimate for written texts since 90-million words (tokens) of the 100- million words of the BNC and BNCweb are from writing. From here these statistics will be our baseline guide.

5. The "387+31 function word" list

Our list of 387 words applied to the BNCweb corpus has the following coverage.

FW387 (tokens)	*Function meaning (-F)*	*Content meaning (-C)*
Function form (F-)	43.3%	1.7%
Content form (C-)	>0.000%	0.3%

5.1. Function form, function meaning (FF)

Of the possible 43.4 percent in the BNCweb these 387 words cover 43.3 percent of FF, just 0.1 percent short of full coverage. In real terms this is a mere 132,043 tokens of the 112,102,325-word BNCweb corpus that is not covered by the list.

5.2. Function form, content meaning (FC)

Of the possible 2.2 percent I have avoided 0.5 percent of FC. It impossible but desirable to avoid rather than to cover these words especially for teaching and learning purposes.

5.3. Content form, function meaning (CF)

Of the four categories CF has the least in quantity at 1 percent. Our list has managed to avoid most of them with coverage at less than 0.000%.

5.4. Content form, content meaning (CC)

Of the 53 percent of words (tokens) which are content words in both form and meaning we have been able to avoid all but 0.3 percent of all instances.

6. Comparing lists

It is all well and good to compare apples with apples when they both come from the same set of data, that is, the BNCweb. So how does this set of 387 function words perform compared to other function word sets. Surprisingly, there are few function words available despite all the peripheral talk within grammar books. We will compare FW387 to three other lists used by O'Shea [O], by Cook [V][xi] and Dang and Webb [D][xii].

The coverage information for each of the lists is

	Function meaning (-F)	*Content meaning (-C)*
Function form (F-)	[O] 40.8% [V] 39.4% [D] 37.0%	[O] 1.9% [V] 1.9% [D] 1.4%
Content form (C-)	[O] 0.0% [V] >0.000% [D] 0.0%	[O] 1.4% [V] 1.3% [D] 0.3%

Of the three lists, the list by O'Shea [O] performed the best of those currently available. In the FF the coverage was still 2.6%[xiii] short in real terms. It also avoided 0.3 percent of FC but showed a large coverage of CC (1.9%). Vivian Cook's list [V] had similar coverage to [O]. Of the 220 word-forms in [V] 172 words were common in both and 48 were different in [V] and 105 were different in [O]. The Dang and Webb list [D] had the worst coverage of FF but, at the same time, it had the best of avoidance of FC as well as CC. Because Dang and Webb's was part of a larger and longer list which included content words avoidance of content words (CC) was possible. However, its coverage of function words (FF) was lacking.

7. Results discussion

Of the lists, my list had the best coverage of function word forms as well as avoidance of content word forms (except for [D] in FC) when compared to the three lists. The 43.3 percent coverage in FF would be hard to improve upon without rendering the list unusable. At 387 words the list is still slim without missing out on all the important words. The 1.7 percent of FC is also reasonable. Only [D] had managed to do better but most likely at the expense of also missing out on important words in FF. The 37 percent level in FF – which is 7.5 percent in real-term non-coverage for [D] – is simply too low. It is also for this reason that [D] managed to do so well in CC. 0.3 percent on paper is a phenomenal achievement on paper. But this is due to the extreme slimness of its list at 176 words.

The main issue with all the compared lists is, firstly, *volume*. At under three-hundred words, the lists were finding it difficult to cover 40 percent of the possible 44.4 percent of function words (FF+CF).

This is especially true of FF. With CF less than 1 percent is possible. Secondly, these lists *must reduce the number of CC*. In attempting to cover CF they have inflated the number of CC. Furthermore, these lists have mistaken content words for function words. Adjectives (a content word) and determiners (a function word), for example, play similar roles in that they modify nouns. Therefore, it is natural to mistake one for the other. Similar examples can be found for other parts-of-speech as well.

Most of the errors in CC by [O] and [V] stem from the inclusion of content words. Of the words from their lists only 'now' and 'somewhat' had function word meaning which accounted for less than 0.1 percent of all instances. In short, the other words found only in [O] and [V] were almost all content words.

8. Conclusion

To properly categorize words is not an easy task. Often, we mistakenly believe form and meaning are one the same. There is always a desire for the simplicity of "one word, one meaning" to thinking about language. This simply neither true nor realistic. The problem starts from the ambiguity of the word "word". Are we talking about the words out in the world being used – the *tokens* of language – or are we talking about in the dictionary as a list of headwords – the *types* of language, when we say "word"? Perhaps we have both these meanings in mind but often we favour one over the other. And sometimes, we choose not to differentiate them. Whatever our reasons, we must make this differentiation.

A word is not *one form, one meaning*. A word form often has multiple meanings. And in between form and meaning is also parts-of-speech. So, the relation chain of form, meaning and parts-of-speech is this:

$$meaning \geq POS \geq form$$

And the ability of POS be categorized into function or content categories make the task of understanding the nature of meaning and form. Of the 44.4 percent of function words (FF+CF) of the BNC/BNCweb this list of 387 word-forms was able to cover 43.3 percent of it. It was also able to avoid most of the content forms with content meaning (CC) at the expense of content forms with function meaning (CF). This is not a great loss since CF only accounts for 1 percent of the total.

While we tend to think that function and content words do not overlap, they in fact do, but only in 3.2 percent of all instances used. Together with the difficulty in differentiating between content and function word POS this overlap tends to be exaggerated or inflated. Hence the importance of hard data over intuition.

But once the hard task of establishing what constitutes a function word (the list) is done then it is only the use of the list is necessary to help make comparisons. Knowing that only 3.2 percent of all function words are ambiguous also helps us visualize and be more confident of being able to say what a function word is.

Function words are important because they make up a large part of the English language – 44.4 percent to be exact. They are also a closed class group of words – between 300 and 500 words depending on how and what one counts. Being so few but covering so much it indicates its importance. To say they have little meaning is to say that we can do without them. But the fact is *we cannot do without them*. Our beliefs and understanding of them are inconsistent. For this single reason alone it is important to reevaluate function words for teaching and understanding.

9. Remaining issues

Because of the nature of cardinal and ordinal numbers, and interjections (totaling 2.3%) as being somewhat an open class group of words, they may perhaps be removed to give a better indication of the true balance of function words. Classifying punctuation (representing 12%) has also been problematic. These have been counted as content words when they are not. An 'undecided' category may help give a better indication of the true numbers of function and content words.[xiv]

References

Anthony, L. (2011). *Laurence Anthony's software*. http://www.antlab.sci.waseda.ac.jp/software.html

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Pearson ESL.

BNCweb. (2012). http://bncweb.lancs.ac.uk/bncwebSignup/user/login.php

Cook, V. (1988). Designing a BASIC Parser for CALL. *Calico*, *6*(1), 50–67.

Dang, T., & Webb, S. (2016). Evaluating lists of high frequency vocabulary. *International Journal of Applied Linguistics, 167* (2), 132–158.

Fries, C. C. (1977). *Structure of English* (New Version). Prentice Hall Press.

Hoffmann, S., Evert, S., Smith, Ni., Lee, D., & Prytz, Y., Berglund. (2008). *Corpus LInguistics with BNCweb—A Practical Guide*. Peter Lang Publishing.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus* (1st ed.). Pearson ESL.

O'Shea, J., Bandar, Z., & Crockett, K. (2012). A Multi-classifier Approach to Dialogue Act Classification Using Function Words. In N. T. Nguyen (Ed.), *Transactions on Computational Collective Intelligence VII* (pp. 119–143). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-32066-8_6

O'Shea, J., Bandar, Z., & Crockett, K. (2011). Using a Slim Function Word Classifier to Recognise Instruction Dialogue Acts. In J. O'Shea, N. T. Nguyen, K. Crockett, R. J. Howlett, & L. C. Jain (Eds.), *Agent and Multi-Agent Systems: Technologies and Applications* (pp. 26–34). Springer Berlin Heidelberg.

Pennebaker, J. (2011). *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language* (2nd Revised Version). Pearson Japan.

Scott, M. (1996). *WordSmith Tools*. http://www.lexically.net/wordsmith/

Appendices (A~L)

A. List of prepositions (81 types)

aboard
about
above
across
after
against
along
alongside
amid
among
amongst
around
as
at
before
behind
below
beneath
beside
besides
between
beyond
but
by
concerning
considering
despite
down
during
except
following
for
from
in
including
inside
into
less
like
minus
near
no
notwithstanding
of
off
on
onto
out
outside
over
past
pending
per
plus
regarding
round
since
that
through
throughout
thru
til
till
to
toward
towards
under
underneath
unlike
until
unto
up
upon
versus
via
vice
vs
with
within
without
worth

B. List of pronouns (64 types)

- anybody
- anyone
- anything
- everybody
- everyone
- everything
- he
- her
- hers
- herself
- him
- himself
- his
- i
- it
- its
- itself
- lots
- me
- mine
- my
- myself
- nobody
- none
- nothing
- one
- oneself
- our
- ours
- ourselves
- plenty
- she
- somebody
- someone
- something
- that
- thee
- their
- theirs
- them
- themselves
- these
- they
- this
- those
- thou
- thy
- us
- we
- what
- whatever
- whatsoever
- which
- whichever
- who
- whoever
- whom
- whose
- ye
- you
- your
- yours
- yourself
- yourselves

C. List of determiners (48 types)

- a
- all
- an
- another
- any
- both
- each
- either
- enough
- every
- few
- fewer
- former
- half
- her
- his
- its
- latter
- less
- little
- many
- me
- more
- most
- much
- my
- neither
- no
- our
- own
- same
- several
- some
- such
- that
- the
- their
- these
- this
- those
- thy
- what
- whatever
- whatsoever
- which
- whichever
- whose
- your

D. List of interjections (49 types)

- aah
- ah
- aha
- alas
- amen
- aye
- blah
- bye
- cheers
- crap
- damn
- dear
- eh
- farewell
- goddamn
- goodbye
- goodnight
- gosh
- ha
- hello
- hey
- hi
- hiya
- hmm
- ho
- huh
- hurray
- hush
- mhm
- nah
- nay
- no
- nope
- oh
- ok
- okay
- ooh
- oops
- ouch
- right
- uh
- um
- urgh
- well
- wow
- yeah
- yep
- yes
- yum

E. List of conjunctions (40 types)

- according
- after
- albeit
- although
- and
- as
- because
- before
- but
- considering
- except
- for

if
immediately
lest
like
nor
notwithstanding
once
or
plus
provided
providing
since
so
supposing
than
that
though
til
till
unless
until
when
where
whereas
whereupon
whether
while
whilst

F. List of ordinal numbers (38 types)

billionth
eighteenth
eighth
eightieth
eleventh
fifteenth
fifth
fiftieth
first
fortieth
fourteenth
fourth
hundredth
last
millionth
nd
next
nineteenth
ninetieth
ninth
rd
second
seventeenth
seventh
seventieth
sixteenth
sixth
sixtieth
st
tenth
th
third
thirteenth
thirtieth
thousandth
trillionth
twelfth
twentieth

G. List of cardinal numbers (33 types)

billion
eight
eighteen
eighty
eleven
fifteen
fifty
five
forty
four
fourteen
hundred
million
nine
nineteen
ninety
one
seven
seventeen
seventy
six
sixteen
sixty
ten
thirteen
thirty
thousand
three
trillion
twelve
twenty
two
zero

H. List of primary auxiliary verbs (32 types)

am
are
aren
be
been
being
d
did
didn
do
does
doesn
doing
don
done
had
hadn
has
hasn
have
haven
having
is
isn
m
re
s
ve
was
wasn
were
weren

I. List of modals (26 types)

can
cannot
could
couldn
d
dare
ll
may
might
mightn
must
mustn
need
needn
ought
oughtn
shall
shalt
shan
should
shouldn
used
will
won
would
wouldn

J. List of adverb-particles (17 types)

about
across
along
around
back
by
down
in
off
on
out
over
round
through
thru
under
up

K. List of WH-adverbs (7 types)

how
however
when
whenever
where
wherever
why

L. Miscellaneous (5 types)

not s t there to

[i] Also called ‘structure words’ and ‘functors’.

[ii] Also called ‘lexical’ words.

[iii] BE, HAVE and DO are have the dual role of primary auxiliary verbs and main verbs. The CLAWS 5 tagset does differentiate between these. While it is best to make differentiation we have followed the original tagset.

[iv] See (Quirk et al., 1985) for a full list. This paper adds adverb particles and WH-words (POS tag AVQ) as well.

[v] This is the standard definition for words as used by concordancers like AntConc (Anthony, 2011). Other concordancers like Wordsmith Tools(Scott, 1996) will count numbers.

[vi] In this paper, morphology and syntax are considered the two main parts of what is called grammar.

[vii] This has already begun with theories like functional grammar and cognitive linguistics.

[viii] The BNCweb is based upon the British National Corpus (BNC), a 100-million token reference corpus of predominantly British English. It consists of 90-million words of written English and 10-million words of spoken English from a variety of genres collected in the 1980s and 1990s. Particularly for this research on function words, there are fewer problems “vocabulary” since function words are stable relative to content words. Function words are less topic-influenced but more genre-influenced.

[ix] See (Leech et al., 2001) and (Hoffmann et al., 2008) for detail studies.

[x] This is not the best solution for both punctuations and unclassified cases and needs to be addressed in future research.

[xi] The letter [V] was chosen in order to avoid confusion with shorthand for content words (C). V is for Vivian, the author’s first name.

[xii] See (Cook, 1988; Dang & Webb, 2016; O’Shea et al., 2012, 2011). Cook’s list was designed for parsing second language student writing. Dang and Webb’s list was part of a larger list – the Essential Word List (EWL) as a learning guide for second language learners. O’Shea’s papers were machine parsing but for natural language processing (NLP) and machine learning.

[xiii] Rounding error.

[xiv] A preliminary study seems to indicate that the percentage for *content forms with content meaning* (CC) is overstated by 13% mostly because of the inclusion of punctuation and unclassified words. This means that the ration of function to content words are roughly about the same at slightly over 40% each, much less than the 60% for function words that is popularly quoted (Pennebaker, 2011).