

Google the verb

Adam Kilgarriff
Lexical Computing Ltd

Abstract

The verb *google* is intriguing for the study of morphology, loanwords, assimilation, language contrast and neologisms. We present data for it for nineteen languages from nine language families.

The Case

There are several reasons why the verb *google* is an appealing object for linguistic research.

- It exists in many languages, with the same core meaning. (For most words it does not make sense to say that the same word exists in many languages. However names, and technical terms, can be language-independent. For *google*, it does seem to make sense to say that the ‘same’ verb exists in many languages.)
- It is new: it has not had time to develop idiosyncratic morphological, phonological or syntactic behaviour, so, like the invented words used in psycholinguistic experiments, it allows us to view the default behaviour for each language
- Unlike invented words, it is common and can be explored using corpus methods
- Most new words are nouns, but verbs tend to show more morphological and syntactic complexity so support a wider range of research questions
- For English, *google* is phonetically and orthographically an unexceptional word which readily adopts standard inflections and other kinds of linguistic variation in speech and in writing. (This does not apply to *Yahoo!*, in speech or in writing). We think this will be fairly true for *google* in at least some other languages, though that is an outcome rather than an input to the research
- As a search term, *google* works well and is easily searched for, in all of its variant forms, in most of the languages we have investigated.

In our corpus query tool, the Sketch Engine, we have general, recent web corpora for a number of languages, gathered as described in Baroni et al 2009, Sharoff 2006, Kilgarriff et al 2009. In the tool we can conveniently search for all forms of the verb, and compute their frequencies-per-million, so, where we had a suitable corpus, this was done. In other cases, a commercial search engine was used.

The Data

Germanic languages

Dutch	NIWaC	128m	English	UKWaC	1,527m
google	1sg, n	670	google	base, n	2488
googlen, googelen, googleen, google-en, goegelen, google'n	inf, 1,2,3 pl, n	55	googling, googleing	prespart, gerund	243
googled, googelt, googlet	2, 3, sg, n	16	googled	past, pastpart	178
googelde, googlede	past sg	2	googles	3 sg, n pl	22
gegoogled, gegooget, gegooget, gegooget, gegooget	pastpart	37	Total	1.98 pm	3031
Total	6.7pm	862	Norwegian	newspaper	788m
German	DeWaC	1,627m	google	infin	259
google, googel, googl, goegele	1 sg	1395	googler	present	99
googlen, googln, googeln, googleln, gugeln	infin, 1,3 pl	681	googlet, googla	past, pastpart	54
gegooglet, gegooget, gegooget, gegooget	pastpart, 3 sg, 2	480	googles	passive	3

gegoogelt	pl		googlede	pastpart def	1
googlet, googled, googelt	3 sg, 2 pl	105	googlende	prespart	1
googelte, googlete	past 1 sg, 3 sg	10	Total	.52 pm	417
googlest, googelst	2 sg	39			
gegoogelte, googelnde	pastpart adj f sg	5	Swedish	informal web	18m
gegoogelten	pastpart adj pl	2	googla	infin	23
ergoogle	1 sg	1	googlar	pres	11
ergooglen, ergoogeln, ergugeln	infin, 1 pl, 3 pl	51	googlade	past	6
ergoogelt, ergooglt, ergooglet	pastpart, 3 sg, 2 pl	51	googlat	supine	13
ergoogelte	past 1 sg, 3 sg	7	googlande	prespart	5
ergoogeltes	pastpart adj neuter	2	Total	3.2 pm	58
ergoogled	3 sg, 2 pl	1			
Total	.315 pm	513			

Notes for data in all tables:

- Inclusion
 - variants for the same item in the verbal paradigm are comma-separated
 - only verb forms included, although counts include nouns as well where the same form can be noun or verb. In these cases the noun option is indicated after semi-colon
 - derivational morphological not included, except where noted below
- order: forms listed in frequency order, or, where that disguises the structure of the paradigm, standard paradigm order
- normalisation: all Latin-alphabet characters normalised to lowercase except where uppercase indicated a name or a noun: then, those cases were excluded
- corpus name is given where this has been used in publications or on the Sketch Engine website; in other cases we give a minimal description of the corpus type, or a note of the search engine used for direct web-searching
- the naming of grammatical roles cannot be done with precision where space is limited and the data covers a wide range of languages, and this is in any case marginal to the paper. Grammatical labels are indicative only. Where no tense is given, tense is present; where no mood is given, mood is indicative. A comma indicates syncretism: the form realises multiple grammatical roles.
- Frequencies per million (for the verb as a whole) are given in most cases where the corpus size is known, in an attempt to make it possible to compare behaviour between languages. However these figures are to be viewed with great caution, not only because the corpora differ in a wide variety of ways, but also because the noun is always far more common than the verb, and in some cases the overall count given will include many noun cases which could not reliably be distinguished from verbal ones.

Dutch and German show a large number of spelling variants. Amongst other things, in Dutch and German spelling the *le* ending is not standard. Some authors have retained it, others have changed it to *el*, others have deleted the *e* altogether, and couple of authors have covered all bases, with an *l* in both possible places: *googleln*. Frequencies for Dutch and English cannot be compared with others because of syncretism between the verb and the much more common noun. The high frequency (per million) in the Swedish corpus, which was collected explicitly to explore informal language, is noteworthy, though based on low numbers.

We have included German *ergooglen*, a derived verb where the prefix means ‘creative process’. This was a common variant on the base verb with an aspectual meaning contrast: see also notes on Slav languages and Chinese below. Other prefixed forms are not included in the table: the second most frequent was *rumgooglen*, a contraction of *herumgooglen* meaning “google around”, which always occurred in collocation with a quantity expression, usually *ein bisschen rumgooglen*, “google around a bit”.

Romance languages

Italian	ItWaC	1,909m	Romanian	Web via Google	
googlare	infin	29	googăli, gugăli	infin	7210
googlato	pastpart	27	googălesc, gugălesc	1 sg, 3 pl	6780
googlando	gerund	26	googălești, gugălești	2 sg	4670
googlate	imper pl, n pl	18	googălește, gugălește	3 sg, imper sg	6500
googla	imper sg, 3 sg	8	googălim, gugălim	1 pl	1387
googlo	1 sg	3	googăliți, gugăliți	2 pl, imper pl	1804
googlò	past	1	googălit, gugălit	pastpart, future	20,430
googlasse	subj, 3 sg	1	googăleam, gugăleam	past cont 1 sg	514
Total	.059 pm	114	googăleai, gugăleai	past cont 2 sg	10
			googălea, gugălea	past cont 3 sg	5
Spanish	Internet Es	117m	googăleați	past cont 2 pl	1
googleando	gerund	11			
googlear	infin	8			
googleo	1 sg	1			
googleas	2 sg	1			
googleadme	imper + pronoun	1			
Total	0.19 pm	22			

In Spanish and many other languages, pronouns are sometimes written attached to the verb, as in *googleadme*, which is included to illustrate the issue and because, after detaching the pronoun, the remaining form is the only imperative found for Spanish.

Slav languages

Czech	Web crawl	800m	Slovak	SNK 4.0	526m
googlen	passive	1	googlovat'	infin	7
progooglovat	"google through" infin	1	googlujú	3 pl	1
progoogluj	"google through" imper	1	googluj	imper 3 sg	1
vygooglovat	"find by google"	1	gúgli	imper 3 sg	1
Total	.005 pm	4	gúglit'	infin imperf	1
			nagooglit'	infin perf	1
Russian	Web crawl	188m	pogooglovat'	infin	1
погуглите	imper pl	6	pregooglujú	3 pl	1
погуглил, нагуглил	past 3 sg m	3	negooglovali	past 3 pl neg	1
погуглила	past 3 sg f	2	vygooglit'	infin perf	2
гуглить	infin imperf	2	vygooglite	2nd pl	1
гуглю	1 sg	2	vygoogli	imper 3 sg	1
погуглить, нагуглить	infin perf	2	vygooglených	pastpart gen pl	1
гуглят	3 pl	1	vygooglené	pastpart nom pl	1
погуглив	past gerund	1	vygooglim	1 sg	1

прогугли	imper sg	1	vygooglovat'	infin	2
Total	.106 pm	20	vygooglujem	1 sg	1
			vygooglovaná	pastpart nom f	1
Slovene	FidaPLUS	620m	vygooglovali	past 3 pl	1
guglanje, googlanje	gerund	8	vygooglovala	past 3 sg f	1
poguglati, pogooglati	infin	7	vygooglujeme	1st pl	1
guglati, googlati	infin	6	v ygúglená	pastpar nom f	1
prigooglati	infin	4	v ygúgli	imper 3 sg	1
Total	.040 pm	25	v ygúglili	past 3 pl	1
			zagúglite	2 pl	1
			Total	.063 pm	33

Amongst the Slav languages we have included verb forms with prefixes relating to aspect. While they are usually treated as derivational morphology, aspect is often conveyed by inflectional and other grammatical means in other language so they have been included here.

We are struck by the very low frequencies for Czech: we wonder if this is because this particular corpus includes more formal data than some others (compare the Swedish, which is informal by design), or because Ceznam, not Google, is the leading search engine in the Czech Republic, or for more linguistic reasons: perhaps Czech is not a language that forms verbs so readily.

Celtic languages

Irish	Web via google	Welsh	Web crawl	120m	
googláil, gúgláil, ghoogláil	gerund	36	gwglo, googlo, googlio, gwglío	base v, n	207
ghoogláil, ghúgláil	infin	25	gwglwyd	impers perf	4
googlóidh	future	2	gwglwch, googlwch	imp pl, 2 pl	2
googlaigh, gúgal	imperative	2	googlia, gwglia	imp sg	2
ghooglaigh	past	1	gwglais	1 sg perf	1
gúgaláilte	verbal adj	1	Total	1.80 pm	216

The Welsh derived forms included *gwglbomio*, 'googlebombing'.

Greek	GkWaC	149m
γκουγκλάρω, γκουγκλίζω	1 sg	17
γκουγκλάρουμε	1 pl	1
γκουγκλάρουν	3 pl	1
googláρεις	2 sg	2
γκούγκλαρα, googlaρα, γκούγκλιζα	past cont, 1 sg	7
γκούγκλιζες	past cont, 2 sg	1
γκούγκλισα	past, 1 sg	5
γκουγκλίσει	subj, 3 sg	1
γκουγκλάροντας, googlίζοντας	gerund	4
γκουγκλίστε	imper, 2 pl	1
γκούγκλισον, ξαναγκούγκλισον	imper, 2 sg	4

Total	.29 pm	44
--------------	---------------	-----------

The variants of the imperative on the last line are formal and a little archaic.

Asian languages

Chinese	Web via baidu		Persian	Web via google	
谷歌一下, google一下	+ aspect	790,000	گوگل می کنم , گوگل میکنم	1 sg	24,710
去谷歌一下, 去google一下		47,400	گوگل می کنم , گوگل میکنی	2 sg	52
可以谷歌一下, 可以google一下		31,463	گوگل می کند, گوگل میکند	3 sg	74,618
上谷歌搜索, 上google搜索		20,400	گوگل می کنیم, گوگل میکنیم	1 pl	71
去谷歌上查一下, 去google上查一下		174	گوگل می کنید, گوگل میکنند	2 pl	67
			گوگل می کنند, گوگل میکنند	3 pl	58,049
Hindi	HindiWaC	34m	گوگل کردن	infinitive	4810
गूगलाया	past	1	گوگل کردم	past 1 sg	3520
गूगले कर	base	1	گوگل کردی	past 2 sg	3370
गूगलाते	"by searching"	1	گوگل کرد	past 3 sg	3160
Total	.088pm	3	گوگل کردیم	past 1 pl	49
Telugu	TeluguWaC	3.4m	گوگل کردید	past 2 pl	140
గూగుల చ్చేసాడు	with light verb	2	گوگل کردند	past 3 pl	960
గూగుల చ్చేసా	light verb, non-finite	4			

The Asian languages covered raise a number of additional issues. Both Persian and Telugu are languages which make extensive and systematic use of light verb constructions, so the verb *google* usually translates as something like the compound verb *do google*.

Chinese has no inflectional morphology and a weaker noun/verb distinction than many languages. It has a writing system without spaces between words and a correspondingly weaker distinction between words and multi-word units. It also presents challenges when one wishes to write a word that one has not seen written before. Aspect markers are the indicators of verb-hood, and here we present the stem (*google* in Latin or 谷歌, the Chinese-writing name of the company) + aspect markers.

In many languages there is an unresolved tension between English-like and localised orthography, applying to, *inter alia*, the choice of character set (in Chinese, Greek) and in the orthographic realisation of the vowel group (with English *oo* not being native to many orthographies: in most cases the alternative is *u*, in Welsh it is *w*.)

Conclusion

We present a data set for the verb *google* across many languages. It presents an interesting testing-ground for a range of ideas on morphology, loanwords, assimilation, language contrast and neologisms. We hope it will stimulate further thinking in these areas.

Acknowledgements

With thanks to Serge Sharoff and the Bologna group for permission to use their corpora in the Sketch Engine. For the specific language expertise I would like to thank: Gisle Andersen, PVS Avinesh, Núria Bel, Vladimir Benko, Sebastian Burghof, Eugenie Giesbrecht, Andrew Hawke, Abhilash Inumella, Håkan Jansson, Vojtěch Kovář, Simon Krek, Monica Macoveiciuc, Mavina Pantazara, Behrang QasemiZadeh, Siva

Reddy, Bettina Richter, Pavel Rychlý, Marina Santini, Simon Smith, Elaine Uí Dhonnchadha, and Carole Tiberius.

References

- Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E. (2009) 'The WaCky wide web: a collection of very large linguistically processed web-crawled corpora'. *Language Resources and Evaluation Journal* 43 (3). 209-226.
- Kilgarriff, A., Reddy, S., Pomikalek, J. (2009) '[Corpus Factory](#).' Proc. Asialex, Bangkok.
- Sharoff, S. (2006). 'Creating general-purpose corpora using automated search engine queries.' In Marco Baroni and Silvia Bernardini, (eds), *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna.