The Sketch Engine as infrastructure for historical corpora

Adam Kilgarriff
Lexical Computing Ltd
Brighton
UK
adam@lexmasterclass.com

Miloš Husák
Lexical Computing Ltd. &
Faculty of Informatics
Masaryk University, Brno, Cz
xhusak@mail.muni.cz

Robyn Woodrow
Lexical Computing Ltd.
Brighton
UK
robyn@sketchengine.co.uk

Abstract

A part of the case for corpus building is always that the corpus will have many users and uses. For that, it must be easy to use. A tool and web service that makes it easy is the Sketch Engine. It is commercial, but this can be advantageous: it means that the costs and maintenance of the service are taken care of. All parties stand to gain: the resource developers both have their resource showcased for no cost, and get to use the resource within the Sketch Engine themselves (often also at no cost). Other users benefit from the functions and features of the Sketch Engine. The tool already plays this role in relation to four historical corpora, three of which are briefly presented.

A premise of historical corpus development is that a corpus, once created, will be widely used. If it is not easy to use it, this will not happen. In 2012, this means making it available to search over the web. You might do this by developing your own tool, or installing and using someone else's, or getting someone else to handle that whole side of things for you.

1 The Sketch Engine

The Sketch Engine is a well-established corpus query tool with a nine-year track record. It is fast, responding immediately for most queries for billion-word corpora, and offers all standard functions (concordancing, sorting and sampling, wordlists, collocates, subcorpora) and some non-standard ones. It takes its name from word sketches, one-page summaries of a word's grammatical and collocational behaviour, as in Fig 1. It is in daily use for lexicography at Oxford University Press, Cambridge University Press, Collins, Cornelsen and Le Robert, for language research at seven national language institutes, and for linguistic and language technology teaching and research at over 100 universities worldwide.

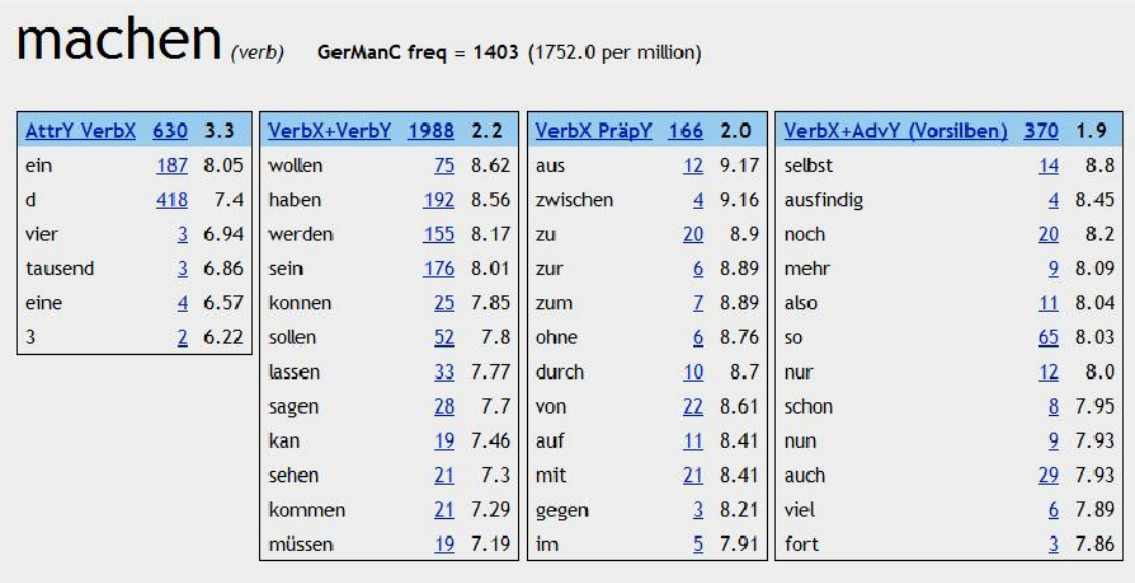
The Sketch Engine is offered as a web service, with 200 corpora for sixty languages already loaded Users may also upload and install their own corpus, and then use the Sketch Engine to study it. Many of the corpora in the tool are provided by their creators, often in exchange for free access for them and their colleagues. The resource developer benefits in three ways:

- access to their own corpus in the Sketch Engine, which supports them in their own research on it (including maintaining and developing it)
- an easy way to show their corpus to others, in a way that allows those others to explore it in detail
- access to other corpora already in the Sketch Engine.

The tool uses input and query formalisms developed at the University of Stuttgart for their corpus system in the early 1990s, as widely adopted across corpus and computational linguistics. There have also been extensions to the formalisms, for example for improved querying of parsed data (Jakubíček et al., 2010).

1.1 Maintenance and motivation

The maintenance of resources has often been a bone of contention for those left in charge of them. Resource developers become the victims of their own success: the more successful the resource, the greater the level of expectation that errors will be corrected and upgrades provided, yet



machen *(verb)* **GerManC freq = 1403** (1752.0 per million)

AttrY VerbX	630	3.3
ein	187	8.05
d	418	7.4
vier	3	6.94
tausend	3	6.86
eine	4	6.57
3	2	6.22

VerbX+VerbY	1988	2.2
wollen	75	8.62
haben	192	8.56
werden	155	8.17
sein	176	8.01
konnen	25	7.85
sollen	52	7.8
lassen	33	7.77
sagen	28	7.7
kan	19	7.46
sehen	21	7.3
kommen	21	7.29
müssen	19	7.19

VerbX PräpY	166	2.0
aus	12	9.17
zwischen	4	9.16
zu	20	8.9
zur	6	8.89
zum	7	8.89
ohne	6	8.76
durch	10	8.7
von	22	8.61
auf	11	8.41
mit	21	8.41
gegen	3	8.21
im	5	7.91

VerbX+AdvY (Vorsilben)	370	1.9
selbst	14	8.8
ausfindig	4	8.45
noch	20	8.2
mehr	9	8.09
also	11	8.04
so	65	8.03
nur	12	8.0
schon	8	7.95
nun	9	7.93
auch	29	7.93
viel	6	7.89
fort	3	7.86

Figure 1: Word sketch for *machen* in the GermanC corpus of early modern German.

research funding bodies are rarely willing to fund them, since the projects have already had their funding, and maintenance is not part of the research funders' mission. So the host organisation struggles to meet users' requests for little credit or recompense. Nor does resource maintenance offer many opportunities to publish.

Lexical Computing, the Sketch Engine company, depends for its income on the quality of its resources, and on users finding the system works well so they renew their licences. It is motivated to maintain and upgrade the hardware, software and corpora. There is an income stream to fund it, from customers.

For resource management and maintenance, there is much to be said for a market model in which the people who are maintaining a resource are motivated to do it well because their income depends on it.

1.2 Local *vs.* remote

One of the biggest questions about software, in the age of the web, is: should it be local or remote? Should we download and install, or interact through browsers and APIs? For a growing number of applications, 'remote' is gaining ground. More and more people manage their documents and photos, and read their email, on remote servers. When I want to convert a document from .ps to .pdf, I do it at http://ps2pdf.com. Corpus research is an area where 'remote' is a very appealing answer, as:

- corpora are large objects which are often awkward to copy
- copying them to other people can be legally problematic
- there are many occasional and non-technical potential corpus users who will not use them if it involves software installation
- the software is more easily maintained and updated
- the user does not need to invest in hardware, or expertise for support and maintenance.

For all of these reasons, the preferred model for most corpus use is the remote one. To support users who want robot access there is a web API.

2 Historical Corpora in Sketch Engine

There are currently historical (pre-20th-century) resources publicly available in the Sketch Engine for three languages: Latin (McGillivray and Kilgarriff, 2011), English and German.[1]

[1]The Sketch Engine is also being used in the ChartEx project (http://www.chartex.org) which is applying text min-

2.1 The Corpus of English Dialogues Corpus

The Corpus of English Dialogues 1560-1760 (Culpeper and Kytö, 2010) was created to explore how English pragmatics developed by gathering historical speech and speech-like data. It comprises 1.2 million words of trial proceedings, witness depositions, play-texts, dialogue in prose fiction and didactic dialogues, including ones from language teaching textbooks. Fig 2 shows a concordance for *prithee*, sorted by date, with the genre of the text also shown. Here we are showing changes of speaker turn by adding the name of the speaker between or-bars, in green and italics (other options are easily set up). Note also the facility for navigating to a particular date.

2.2 Penn Historical Corpora (PHC)

The Penn Historical Corpora are the Penn-Helsinki Parsed Corpus of Middle English (second edition; PPCME2), the Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME), and the Penn Parsed Corpus of Modern British English (PPCMBE). They all comprise texts and text samples of British English prose from the earliest Middle English documents up to the First World War.[2] Fig 3 shows the ascent of *should* over the past 500 years.

A business issue arose when a user asked if they could access the PHC in the Sketch Engine. Penn have been selling the PHC, on CD-Rom, and this has been funding ongoing research and maintenance. So its creator was keen to make the PHC available in the Sketch Engine - but not at the expense of the income stream. The solution we have adopted is that only those users who have bought the CD-Rom will get access to the PHC in the Sketch Engine, and purchasers of the CD-Rom will receive a year's free access to the Sketch Engine so they can look at PHC (and all the other corpora) there.

2.3 GermanC

GermanC (Durrell et al., 2007) is a corpus of 800,000 words of 17th and 18th century German.

We demonstrate what can be done with GermanC in the Sketch Engine by looking at 'keyword lists', the words with the biggest contrast between frequencies in one corpus (or subcorpus) and another. Here we focus on the more frequent words (by adjusting the 'simplemaths parameter' (Kilgarriff, 2009)).

The fifty top mid-to-high frequency lemmas[3] of the 17th century subcorpus of GermanC, in contrast to the 18th century part, include:

> ach allhier also Artikel auch begehren berichten Christus damit dann darauf derselbige dito etlich Feind Fürst gar Gnade Gott halten jenige Kapitel Komet König Leib lieb mit mögen oder Ort Pferd Rat solch sollen sonder sonderlich sonst statt Tod Türke und vom wann weil wider wiederum wohl Zeche

The corresponding 18th century items are:

> Absicht Art Begriff besonders denken derjenige dies eben ein Erde finden Freund für Gegend Gegenstand Geschlecht Graf hier ich immer jeder klein können Körper machen Mann mein Mercurius Mutter Natur nur nötig scheinen schon Seite Sie suchen Teil Ton um Umstand Vater verschieden wahr weit wenigstens wenn wirklich zeigen

The word with the highest frequency contrast, with over double the relative frequency in 17c *vs.* 18c, was *wann*. At the top of the 18c list was *wenn*.[4] Both words are of similar frequency. This strongly suggests that people were making a choice between *wann* and *wenn*, and in the 17th century they more often chose *wann*, and in the 18th, *wenn*. While the changes affecting these two words are already familiar (Wright, 1907; Abraham, 1978), with GermanC in the Sketch Engine we can directly explore exactly what the changes are and when and how they took place.

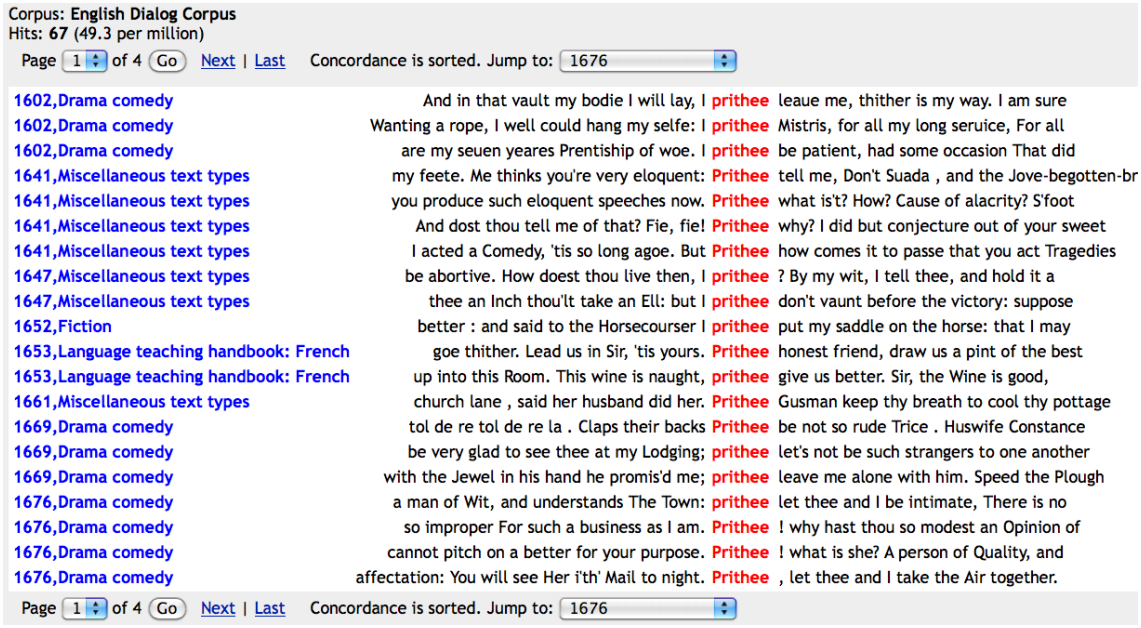
Several other words in the 17c list (*allhier, derselbige*) are marked in dictionaries as old, or obsolete.

ing methods to medieval Latin charters. It will make the corpora it prepares publicly available through the Sketch Engine as the project proceeds.

[2] `http://www.ling.upenn.edu/hist-corpora/`

[3] Lemmatisation was by a version of TreeTagger trained on GermanC data (Scheible et al., 2012).

[4] Here the lists are alphabetisicised. In the tool, *wann* and *wenn* appear at the tops of the two lists.



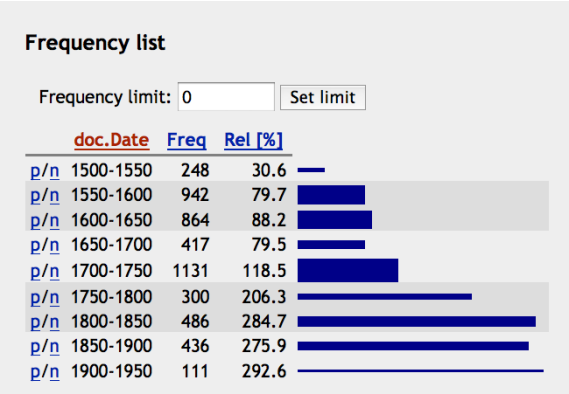
Corpus: **English Dialog Corpus**
Hits: **67** (49.3 per million)

Page 1 of 4 Go Next | Last Concordance is sorted. Jump to: 1676

1602,Drama comedy	And in that vault my bodie I will lay, I	prithee	leaue me, thither is my way. I am sure
1602,Drama comedy	Wanting a rope, I well could hang my selfe: I	prithee	Mistris, for all my long seruice, For all
1602,Drama comedy	are my seuen yeares Prentiship of woe. I	prithee	be patient, had some occasion That did
1641,Miscellaneous text types	my feete. Me thinks you're very eloquent:	Prithee	tell me, Don't Suada , and the Jove-begotten-br
1641,Miscellaneous text types	you produce such eloquent speeches now.	Prithee	what is't? How? Cause of alacrity? S'foot
1641,Miscellaneous text types	And dost thou tell me of that? Fie, fie!	Prithee	why? I did but conjecture out of your sweet
1641,Miscellaneous text types	I acted a Comedy, 'tis so long agoe. But	Prithee	how comes it to passe that you act Tragedies
1647,Miscellaneous text types	be abortive. How doest thou live then, I	prithee	? By my wit, I tell thee, and hold it a
1647,Miscellaneous text types	thee an Inch thou'lt take an Ell: but I	prithee	don't vaunt before the victory: suppose
1652,Fiction	better : and said to the Horsecourser I	prithee	put my saddle on the horse: that I may
1653,Language teaching handbook: French	goe thither. Lead us in Sir, 'tis yours.	Prithee	honest friend, draw us a pint of the best
1653,Language teaching handbook: French	up into this Room. This wine is naught,	prithee	give us better. Sir, the Wine is good,
1661,Miscellaneous text types	church lane , said her husband did her.	Prithee	Gusman keep thy breath to cool thy pottage
1669,Drama comedy	tol de re tol de re la . Claps their backs	Prithee	be not so rude Trice . Huswife Constance
1669,Drama comedy	be very glad to see thee at my Lodging;	prithee	let's not be such strangers to one another
1669,Drama comedy	with the Jewel in his hand he promis'd me;	prithee	leave me alone with him. Speed the Plough
1676,Drama comedy	a man of Wit, and understands The Town:	prithee	let thee and I be intimate, There is no
1676,Drama comedy	so improper For such a business as I am.	Prithee	! why hast thou so modest an Opinion of
1676,Drama comedy	cannot pitch on a better for your purpose.	Prithee	! what is she? A person of Quality, and
1676,Drama comedy	affectation: You will see Her i'th' Mail to night.	Prithee	, let thee and I take the Air together.

Page 1 of 4 Go Next | Last Concordance is sorted. Jump to: 1676

Figure 2: Concordance for *prithee* in the English Dialogues corpus, sorted by date, showing genre and speaker-turns.



Frequency list

Frequency limit: 0 Set limit

	doc.Date	Freq	Rel [%]
p/n	1500-1550	248	30.6
p/n	1550-1600	942	79.7
p/n	1600-1650	864	88.2
p/n	1650-1700	417	79.5
p/n	1700-1750	1131	118.5
p/n	1750-1800	300	206.3
p/n	1800-1850	486	284.7
p/n	1850-1900	436	275.9
p/n	1900-1950	111	292.6

Figure 3: Analysis by time period of *should* in Penn Historical Corpora. Bars show frequency normalised according to the quantity of data available for each time period.

Other notable contrasts are that 17c has more formal texts, talking about religion and royalty, whereas 18c talks more about ordinary life: mothers and fathers and friends. Whether this is an artefact of the selection of texts to go into the corpus, or a reflection of changes in the language or of the role of writing in society, would require investigation beyond the scope of this paper.

An advantage of working in the Sketch Engine is that there are many corpora available for comparison. So we can also make a keyword list for all of GermanC, in contrast to a different corpus, for example deWaC, a web-crawled corpus of contemporary German (Baroni et al., 2009).[5] The keywords of GermanC are:

> aber alle allein also anderer bald da daß dein derselbe doch du eben einig er gar gemacht geschehen gewiß gleich Gott halten Herr Herz ich Ihr König lassen man mein mögen nichts nun Ort Sache sehen Sie so solch sollen sonst tun unser wann was weil welch wo wohl woollen

While some items (*alle, einig, er, Ihr, Sie*) appear in the list owing to lemmatisation differences and others (*daß*) owing to spelling differences, others are linguistic, owing either to the differences in texts included, or to some other differences in German society or language over the last three centuries. GermanC makes far more use of first and second person pronouns, short time adverbials (*bald, doch*), and some conjunctions (*aber, da*). These seem to be indicative of GermanC being, overall, a corpus of less formal texts than DeWaC.

All three lists contain a number of pronouns. We find pronouns at the top of keyword lists time and time again. Pronouns are the litmus paper of text type.

We may suspect that the 18c-17c comparison is very different to the GermanC-DeWaC comparison since the components of GermanC will be, overall, much more similar to each other than GermanC is to DeWaC. For ways to explore this topic see (Kilgarriff, 2001; Kilgarriff, 2012).

[5] As different TreeTagger models were used for the two corpora, there will be slight differences in lemmatisation. With this in mind we also explored the keyword list of word forms. But this was dominated by spelling variants, which had been addressed by giving the normalised form of the lemma, so the lemma list was more informative.

3 Further publishing possibilities

A scenario currently under discussion with one corpus developer involves the Sketch Engine taking on a publisher role, including collecting payments and passing them on to the developer. While some universities have departments that could undertake this role, their costs are often high, and there are benefits to working with a flexible small company with expertise –technical, commercial and legal– in corpora.

4 Conclusion

Two problems often confronting corpus developers are:

1. how to make it easy for everyone to use the corpus
2. how to maintain it and continue to make it available, over a number of years.

We have shown one solution to these problems: subcontract to a commercial company with appropriate tools and expertise. We have shown how this works in several cases, showing a range of the functions and display options that the Sketch Engine offers that are of particular relevance for historical data, and demonstrating how we can immediately make interesting findings using the Sketch Engine.

Acknowledgements

With thanks to Laura Giacomini for assistance with the analysis for German and to Ravi Kiran for the encoding of the Corpus of English Dialogues.

References

Werner Abraham. 1978. The role of fallacies in the diachrony of sentence connectives. *Studies in Second Language Acquisition*, 1(1):95–134.

Marco Baroni, Silva Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Jonathan Culpeper and Merja Kytö. 2010. *Early Modern English dialogues: spoken interaction as writing*. Cambridge University Press.

Martin Durrell, Astrid Ensslin, and Paul Bennett. 2007. The GerManC project. *Sprache und Datenverarbeitung*, 31:71–80.

Miloš Jakubíček, Adam. Kilgarriff, Diana McCarthy, and Pavel Rychlý. 2010. Fast syntactic searching in very large corpora for many languages. In *Proceedings of PACLIC 24*, pages 741–747.

Adam Kilgarriff. 2001. Comparing corpora. *International journal of corpus linguistics*, 6(1):263–276.

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proc. Corpus Linguistics*.

Adam Kilgarriff. 2012. Getting to know your corpus. In *Text, Speech and Dialogue*. Springer.

Barbara McGillivray and Adam Kilgarriff. 2011. Tools for historical corpus research, and a corpus of latin. *New Methods in Historical Corpora*.

S. Scheible, R.J. Whitt, M. Durrell, and P. Bennett. 2012. GATEtoGerManC: A GATE-based Annotation Pipeline for Historical German. *LREC*.

Joseph Wright. 1907. *Historical German Grammar*, volume 1. OUP.