Vietnamese Word Sketches

Adam Kilgarriff
Lexical Computing Limited
UK

Phuong Le-Hong
VNU Hanoi University of Science
Vietnam

Abstract

Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour. They were first used in the production of the Macmillan English Dictionary (Rundell 2002). At that point, word sketches only existed for English. Today, the Sketch Engine is available, a corpus tool which takes as input a corpus of any language and corresponding grammar patterns and which generates word sketches for the words of that language. It also automatically generates a thesaurus and 'sketch differences', which specify similarities and differences between near-synonyms. A web corpus of Vietnamese was tokenized, part-of-speech-tagged and loaded into the Sketch Engine. The results show that word sketches could significantly facilitate lexicographic work for Vietnamese, as they have for other languages. The word sketches, for Vietnamese and many other languages, can be seen at http://www.sketchengine.co.uk.

Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour. Their value for lexicographic work in English and other languages, as well as the background of the use of corpora in lexicography, have been described elsewhere (Kilgarriff and Rundell 2002, Kilgarriff et al. 2004).

A variety of corpus query systems (CQSs) have been used to examine corpus evidence since the rise of the first electronic corpora. Starting with the ground-breaking COBUILD project, lexicographers have been using KWIC (Key Word In Context) concordances as their primary tool for finding out how a word behaves. Later, with the growth of corpora, lexical statistics had to be applied to manage the abundant data and highlight the most salient collocations. Today, state-of-the-art CQSs allow the lexicographer great flexibility in searching for phrases, collocates, grammatical patterns, sorting concordances according to a wide range of criteria, selecting subcorpora for searching in, say, only spoken text, academic text, or only fiction. Available systems include WordSmith (Scott 2008) and the Stuttgart Corpus Workbench (CWB, Christ and Schulze 1994), and the Sketch Engine.

The Sketch Engine (http://www.sketchengine.co.uk) is a corpus query system which gives access to the familiar CQS functions: concordances for several types of queries (simple, lemma, phrase, word form and CQL), with an integrated context control filter (Fig. 1).

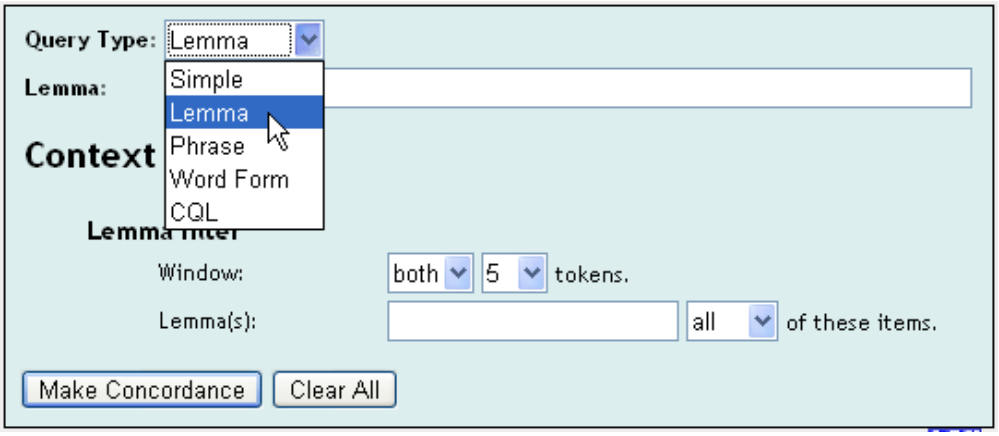
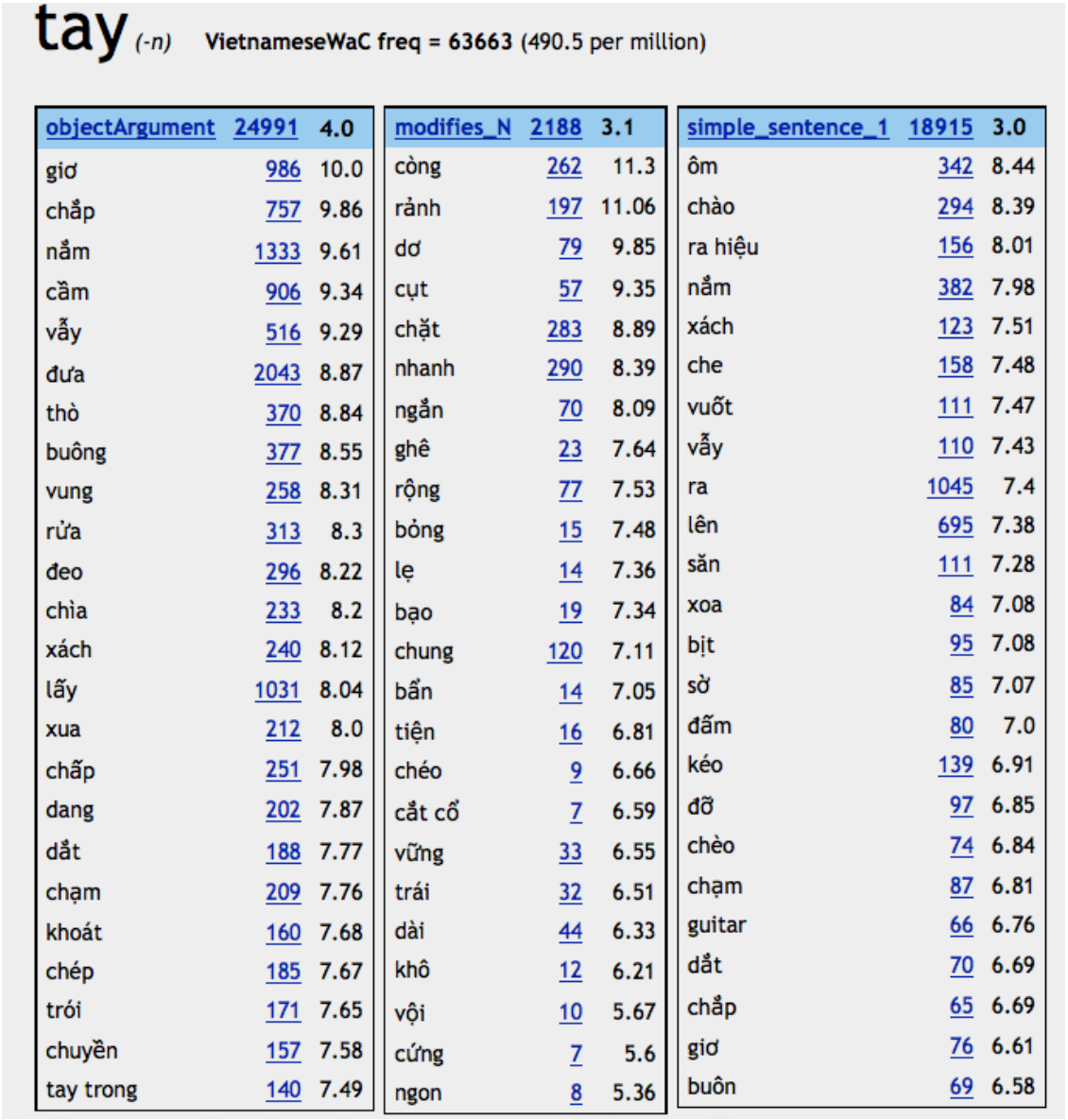


Fig. 1: The Sketch Engine concordancer query window

The features of the Sketch Engine which are of special interest in this paper are not part of standard concordancing programs. These features include word sketches, sketch differences and a thesaurus. They are all fully integrated with standard concordancing.

A word sketch is a one-page automatic, corpus-based summary of a word's grammatical and collocational behaviour. A word sketch for Vietnamese *tay* ('hand') is given in Fig. 2.



tay *(-n)* **VietnameseWaC freq = 63663** (490.5 per million)

objectArgument	24991	4.0
giơ	986	10.0
chắp	757	9.86
nắm	1333	9.61
cầm	906	9.34
vẫy	516	9.29
đưa	2043	8.87
thò	370	8.84
buông	377	8.55
vung	258	8.31
rửa	313	8.3
đeo	296	8.22
chìa	233	8.2
xách	240	8.12
lấy	1031	8.04
xua	212	8.0
chấp	251	7.98
dang	202	7.87
dắt	188	7.77
chạm	209	7.76
khoát	160	7.68
chép	185	7.67
trói	171	7.65
chuyền	157	7.58
tay trong	140	7.49

modifies_N	2188	3.1
còng	262	11.3
rảnh	197	11.06
dơ	79	9.85
cụt	57	9.35
chặt	283	8.89
nhanh	290	8.39
ngắn	70	8.09
ghê	23	7.64
rộng	77	7.53
bỏng	15	7.48
lẹ	14	7.36
bạo	19	7.34
chung	120	7.11
bẩn	14	7.05
tiện	16	6.81
chéo	9	6.66
cắt cổ	7	6.59
vững	33	6.55
trái	32	6.51
dài	44	6.33
khô	12	6.21
vội	10	5.67
cứng	7	5.6
ngon	8	5.36

simple_sentence_1	18915	3.0
ôm	342	8.44
chào	294	8.39
ra hiệu	156	8.01
nắm	382	7.98
xách	123	7.51
che	158	7.48
vuốt	111	7.47
vẫy	110	7.43
ra	1045	7.4
lên	695	7.38
săn	111	7.28
xoa	84	7.08
bịt	95	7.08
sờ	85	7.07
đấm	80	7.0
kéo	139	6.91
đỡ	97	6.85
chèo	74	6.84
chạm	87	6.81
guitar	66	6.76
dắt	70	6.69
chắp	65	6.69
giơ	76	6.61
buôn	69	6.58

Fig. 2. Word sketch for Vietnamese *tay*.

To create the word sketch, the Sketch Engine needs to know how to identify words connected by a grammatical relation. This can be achieved in one of two ways.

The first possibility is to parse the input corpus, so that the information about which word-instances stand in which grammatical relations with which other word-instances is embedded in the corpus. Currently, dependency-based syntactically annotated corpora are supported. We need to mark heads of phrases in phrase structure trees.

In the second approach, the input corpus is loaded into the Sketch Engine part-of-speech-tagged but not parsed, and the Sketch Engine supports the process of identifying grammatical relation instances through a *sketch grammar*. Grammatical relations (or *gramrels*, for short) will be defined one by one, using the Sketch Engine to test and develop them. Once the developer is satisfied with the definition of each grammatical relation, they save the file and the Sketch Engine then compiles it, finding all instances of all grammatical relations in the corpus. It puts them in a gramrels database, which gives users access to word sketches.

Grammatical relations are defined as regular expressions over part-of-speech (POS) tags, using the CQL formalism as first specified within the Stuttgart Corpus Wordbench and extended in Jakubíček et al (2010).

The Corpus

The Vietnamese sketch grammar was developed and tested on a corpus of Vietnamese. The corpus was gathered from the web as described in Kilgarriff et al (2010). It contains 94 million words.

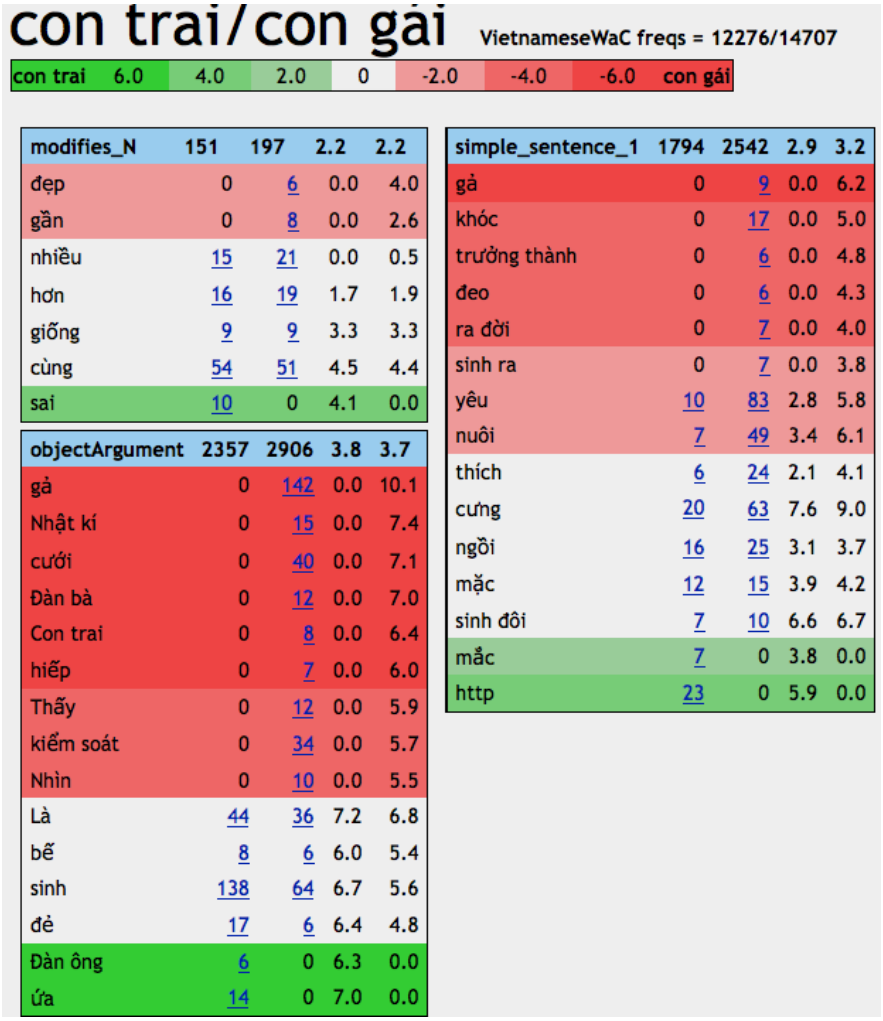
For the corpus to be loaded into the Sketch Engine, it must be tokenized and POS-tagged beforehand. Tokenizing Vietnamese is the task of deciding which space-separated items go together to form words, since words are the objects we need to identify for further processing, and for which we want to write dictionary entries. POS-tagging is the task of deciding the correct word class for each word in the corpus; e.g. determining which words are nouns, which words are verbs, etc. A tagger presupposes a linguistic analysis of the language which has given rise to a set of the morpho-syntactic categories of the language (a tagset).

The Vietnamese web corpus was tokenized and tagged using the tools described in Le Hong (2010).

Word sketches are available for all nouns, verbs, adjectives and adverbs of Vietnamese where there is sufficient data: around ten thousand of the core words of the language.

Sketch differences

Synonyms (and antonyms) tend to share some of the collocates but not all. The *sketch differences* module in the Sketch Engine highlights the shared and different collocational context of two specified words. The listing is colour-coded to show at a glance the commonalities and differences between the lemmas.



con trai / con gái VietnameseWaC freqs = 12276/14707

con trai 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 con gái

modifies_N	151	197	2.2	2.2
đẹp	0	6	0.0	4.0
gần	0	8	0.0	2.6
nhiều	15	21	0.0	0.5
hơn	16	19	1.7	1.9
giống	9	9	3.3	3.3
cùng	54	51	4.5	4.4
sai	10	0	4.1	0.0

objectArgument	2357	2906	3.8	3.7
gả	0	142	0.0	10.1
Nhật kí	0	15	0.0	7.4
cưới	0	40	0.0	7.1
Đàn bà	0	12	0.0	7.0
Con trai	0	8	0.0	6.4
hiếp	0	7	0.0	6.0
Thấy	0	12	0.0	5.9
kiểm soát	0	34	0.0	5.7
Nhìn	0	10	0.0	5.5
Là	44	36	7.2	6.8
bế	8	6	6.0	5.4
sinh	138	64	6.7	5.6
đẻ	17	6	6.4	4.8
Đàn ông	6	0	6.3	0.0
ứa	14	0	7.0	0.0

simple_sentence_1	1794	2542	2.9	3.2
gả	0	9	0.0	6.2
khóc	0	17	0.0	5.0
trưởng thành	0	6	0.0	4.8
đeo	0	6	0.0	4.3
ra đời	0	7	0.0	4.0
sinh ra	0	7	0.0	3.8
yêu	10	83	2.8	5.8
nuôi	7	49	3.4	6.1
thích	6	24	2.1	4.1
cưng	20	63	7.6	9.0
ngồi	16	25	3.1	3.7
mặc	12	15	3.9	4.2
sinh đôi	7	10	6.6	6.7
mắc	7	0	3.8	0.0
http	23	0	5.9	0.0

Fig. 3: Sketch difference: *con trai* versus *con gái*

Our example in 3 shows a differential profile for two closely related V, equivalents to English *boy* and *girl*. The nouns in red (in the web interface; shaded and towards the top of each column, in the greyscale screenshot) exhibit collocational preference to *con trai*, whereas those in green (and towards the bottom of each column) to *con gái.* Items against the white background collocate well with both adjectives. Sketch

differences address the teasing apart of near-synonyms, one of the more difficult aspects of language description and use, as well as (as here) contrasting the collocational patterns of related words.

Conclusion and further work

We have presented a sketch grammar for Vietnamese and shown how, through the Sketch Engine, it can be used to explore the grammar and lexis of Vietnamese, at http://www.sketchengine.co.uk. The grammar is far from perfect and the project is ongoing, with short-term goals including a review of tokenisation. As the output of the Sketch Engine is only as good as the underlying corpus, another agenda item is the preparation of an improved, extended corpus. Despite these limitations, we believe that Vietnamese word sketches are already a resource that is well able to support Vietnamese lexicography and linguistic research in ways that have not been possible before.

References

Christ, O. and M. Schulze. (1994). The IMS Corpus Workbench: Corpus Query Processor (CQP) User's Manual University of Stuttgart. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

Jakubíček, M., A. Kilgarriff, D. McCarthy and P. Rychlý (2010). Syntactic searching in very large corpora for many languages. In: Otoguru, R. et al. (eds.). *Proceedings of Workshop on Advanced Corpus Solutions, PACLIC 24*.

Kilgarriff, A. and M. Rundell (2002). Lexical profiling software and its lexicographic applications - a case study. In: Braasch, A. and C. Povlsen (eds.). *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002*. Copenhagen: Center for Sprogteknologi, Copenhagen University. 807–818.

Kilgarriff, A., P. Rychlý, P. Smrž and D. Tugwell (2004). The Sketch Engine. In: Williams, G. and S. Vessier (eds.). *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004*. Lorient: Université De Bretagne Sud. 105–116.

Kilgarriff, A., S. Reddy, J. Pomikalek, Avinesh PVS (2010). A Corpus Factory for Many Languages. LREC, Malta.

Le-Hong Phuong (2010). Elaboration d'un composant syntaxique à base de grammaires d'arbres adjoints pour le vietnamien. PhD Thesis, Univ. Nancy 2, France.

Rundell, M. (ed) (2002). *Macmillan English Dictionary for Advanced Learners*. Macmillan Education.

Scott, M. (2008). WordSmith Tools version 5, Liverpool: Lexical Analysis Software.