D4.1: Methodology of Sketch Grammar evaluation

The evaluation of word sketch grammars can be performed in a number of ways depending on the particular purpose word sketches are used for. In the following we however focus on a general-purpose approach to the word sketch evaluation and related issues.

Quantitative evaluation

Quantitative evaluation of word sketches can be useful especially when developing a new sketch grammar or trying to improve the coverage of an existing one. The quantitative evaluation in this case consists merely of estimating what percentage of corpus tokens is covered as headwords with meaningful grammatical relations.

This information is useful for potential improvements of individual grammatical relations as defined using the CQL queries (Jakubicek, 2010). The Sketch Engine (Kilgarriff, 2004, 2014) has been recently enhanced by feature that directly shows the proportion of corpus positions covered by the displayed word sketch table for a given headword.

test *(noun)* Alternative PoS: verb (6,896)
British National Corpus freq = 15,789 (140.74 per million) Coverage: 92.85%

modifier	8,554	1.80
driving	127	8.64
blood	204	8.19
fitness	82	7.93
beta	69	7.91
u	49	7.52
diagnostic	50	7.44
screening	46	7.22
means	97	7.15
breath	75	7.14
nuclear	75	7.10
exact	43	7.04
laboratory	62	7.01
statistical	42	6.96
liver	42	6.96
acid	70	6.94
Hiv	39	6.92
written	45	6.91
intelligence	52	6.91
rank	51	6.86
antibody	39	6.81
reasonableness	31	6.80
forensic	31	6.80
smear	31	6.80
positive	58	6.78
iq	29	6.75

modifies	3,517	0.90
cricket	105	8.65
match	170	8.16
statistic	32	8.07
kit	60	7.95
ban	58	7.83
score	76	7.81
tube	60	7.79
wicket	35	7.71
pilot	56	7.50
debut	33	7.18
cricketer	20	7.17
meal	60	7.13
certificate	40	6.92
suite	22	6.76
series	85	6.59
item	61	6.52
innings	12	6.44
result	115	6.25
rig	10	6.16
treaty	30	6.16
strip	17	6.12
igg	8	6.10
batsman	10	6.10
flight	29	6.07
drive	27	6.02

object_of	3,461	2.10
pass	294	8.29
undergo	53	7.78
satisfy	49	7.75
fail	134	7.55
devise	39	7.39
conduct	69	7.39
administer	31	7.24
perform	77	7.23
apply	101	6.82
boycott	13	6.68
carry	120	6.32
stand	80	6.24
standardise	8	6.11
withstand	9	6.10
face	57	6.06
screen	10	5.99
use	227	5.44
repeat	16	5.41
undertake	18	5.37
survive	13	5.36
introduce	36	5.35
play	66	5.33
score	12	5.32
develop	48	5.27
win	37	5.12

and/or	1,476	0.70
international	13	7.58
histology	7	7.09
x ray	11	7.09
examination	28	6.23
exam	8	6.17
Southampton	6	6.12
board	42	5.34
interview	12	5.11
check	6	5.04
measurement	6	4.86
trial	13	4.81
blood	12	4.64
assessment	11	4.57
laboratory	6	4.50
curriculum	6	4.25
procedure	12	4.16
exercise	7	4.00
analysis	13	3.92
result	21	3.85
treatment	10	3.70
culture	8	3.69
datum	6	2.76
patient	7	2.59
case	10	2.10
study	6	1.73

subject_of	1,123	1.20
reveal	51	6.48
match	19	6.14
prove	44	6.00
consist	13	5.42
detect	9	5.41
show	124	5.37
confirm	17	5.23
measure	13	5.17
check	13	5.09
assess	7	4.34
determine	14	4.28
result	8	4.28
apply	15	4.19
demonstrate	6	4.05
run	25	4.00
involve	20	3.98
fail	10	3.96
drive	6	3.31
indicate	6	3.25
identify	6	3.13
provide	24	3.06
lay	6	2.96
follow	16	2.96
remain	8	2.89
suggest	9	2.87

Figure 1: Sample word sketch table for the English noun *test* with its coverage displayed.

In Figure 1 the screenshot from Sketch Engine for the word sketch of the English noun *test* shows a coverage of 92.85 %, meaning that this percentage of all occurrences of *test* as noun in the corpus is covered by some grammatical relation defined in the word sketch grammar and hence the collocations are captured by the word sketch table as presented.

Qualitative evaluation

Extrinsic evaluation

The qualitative evaluation of word sketches is of course a much more complex task -- as it is usually the case. Given the practical nature of word sketches, ideally they should be evaluated extrinsically by measuring their contribution to improvements given a particular task.

A candidate task would be a real lexicographic setting where e.g. one group of lexicographers would be working on dictionary entries while using Sketch Engine with sketch grammar A and another group would be using the same setup with sketch grammar B. One could evaluate the group performance in terms of speed but also in terms of quality of the entries, number of examples etc.

Such evaluations have been performed as in-house trials by major UK publishing houses (e.g. Oxford University Press, Cambridge University Press, Macmillan etc.) who are using Sketch Engine but only in very small scale and only to compare with a null variant -- i.e. not using Sketch Engine.

Intrinsic evaluation

First experiments with both quantitative and qualitative evaluation of word sketches have been performed in 2010 for Dutch, English, Japanese and Slovene, Japanese (see Kilgarriff et al. 2010). In these experiments only precision of the collocations automatically found by word sketches was evaluated, not the recall. Therefore, one was able to answer the question: *how many of the automatically found collocations are good?* but not *which good collocates are missing?* For this some form of a gold standard of collocations for a set of headwords would be necessary so as to have a reference set to be compared with.

In 2011 we therefore setup (and published in Kilgarriff, 2012) an intrinsic evaluation task we where have collected a gold standard of collocates (disregarding which grammatical relation they would fit in) by hiring a group of professional lexicographers (for English) and linguist students (for Czech) as annotators, providing them with collocation candidates and asking the question *Would you put this word into a collocation dictionary?* The answers were binary yes or no.

Of course this very much depends on the actual definition of a collocation (which later turned out to be the biggest methodological problem). The definition that the annotators were given was the one given in the Oxford Collocations Dictionary:

Collocation is the way words combine in a language to produce natural-sounding speech and writing. ... Combinations of words in a language can be ranged on a cline from the totally free — see a man/car/book — to the totally fixed and idiomatic — not see the wood for the trees. ... All these combinations, apart from those at the very extremes of the cline, can be called collocation. And it is combinations such as these — particularly in the 'medium-strength' area — that are vital to communicative competence in English. (Crowther et al., 2002, vii)

When preparing the collocation candidates we collected a number of text corpora (both web-based and edited-content-based) and existing collocations dictionaries. The headwords were only nouns, adjectives and verbs chosen randomly from three frequency bands:

- high: top 100–2999 words by frequency
- mid: top 3000–9999 words by frequency
- low top 10,000–30,000 words by frequency

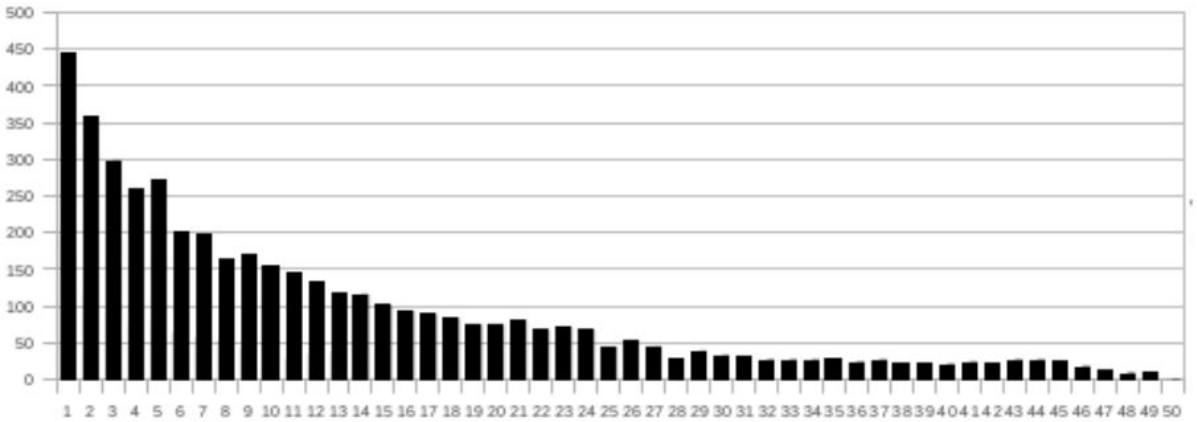

Figure 2: Distribution of good collocations in fiftieths, ordered by score.

We used various sources to achieve close to 100 % coverage for the collocation candidates and we clearly showed that there is a strong correlation between the word sketch score and the goodness of a collocation (as provided in Figure 2) nevertheless our later experiments have shown that a full coverage was not achieved, which of course represents a methodological issue from the point of view of using such a dataset as a gold standard. With respect to this, the findings were blatant in saying:

• *if we showed judges more candidates from the same corpora, they found more collocations (though with diminishing returns)*

• *if we showed judges more candidates from new corpora,*
they found more collocations.

It is worth noting that the task as presented in (Kilgarriff, 2014) was setup an extrinsic corpus evaluation using word sketches -- of course, this becomes then intrinsic evaluation from the perspective of a word sketch grammar.

The underlying assumption was that, having a processing pipeline for a corpus as shown in Figure 3, one can evaluate any of the components by interchanging it but keeping the rest the same.

Figure 3: sample processing pipeline from corpus source to collocations

We have evaluated a large number of corpora with different processing tools and various parameters -- including different taggers and word sketch grammar exploiting their annotation.

The most cumbersome issue remained the definition of collocation which materialized in rather low inter-annotator agreement: for Czech, the pairwise agreement varied between 73.6% and 82.3%, for English, between 81.1% and 85.8% (with three annotator for both languages). Later investigations revealed that for some of the headwords (such as *plutonium*), no pair of annotators agreed on a good collocation. We have found out that the guidance to the annotators was clearly adopted with a printed dictionary on mind, where the set of collocations is rather small and focuses only on very strong collocations.

However, what we were actually most interested in, was an evaluation which would clearly distinguish between obvious errors (i.e. non-collocations) and remaining mild, medium or strong collocations. Such type of evaluation could be used to improvements alongside the whole processing pipeline, including word sketch grammar, in a rather straightforward way. In some cases, we have also noticed that, since the words were chosen randomly, there were terms from domains which the lexicographers were not familiar with, and hence were unable to identify good and strong collocations.

Recently, we have therefore reviewed the methodology and we are now in preparation of a new gold standard collocation set following a revised methodology for annotation that aims to be more inclusive. The annotators are now classifying into five categories:

- strong collocation
- weak collocation
- correct word combination but not a significant collocation
- error
- I don’t understand

To help reducing the number of unknown collocations, the word sketches have been enhanced by the so called longest-commonest match (LCM) string -- the most common headword-collocation combination (Kilgarriff et al. 2015).

We have also devised a detailed annotation manual that would help the annotators to decide among these categories which is provided in Figure 4:

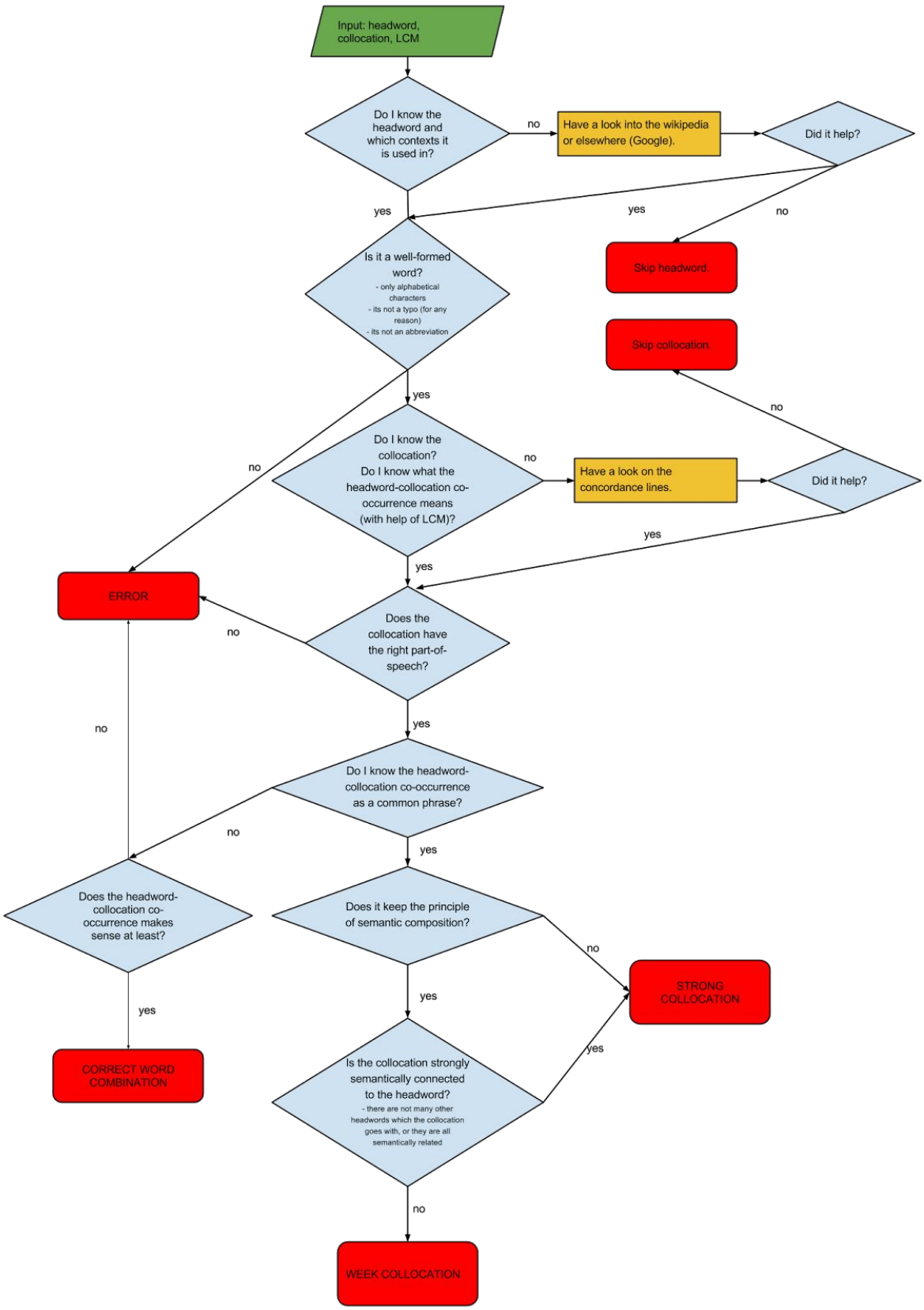

Figure 4: New annotation manual for collocations classification

References

JAKUBÍČEK, Miloš, et al. Fast Syntactic Searching in Very Large Corpora for Many Languages. In:PACLIC. 2010. p. 741-747.

KILGARRIFF, Adam, et al. A quantitative evaluation of word sketches. In: *Proceedings of the 14th EURALEX International Congress, Leeuwarden, The Netherlands*. 2010.

KILGARRIFF, Adam, et al. Extrinsic corpus evaluation with a collocation dictionary task. In:*Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. 2014.

KILGARRIFF, Adam, et al. Longest–commonest Match. 2015.