



Dominic Stewart*

USING WORD SKETCH TO INVESTIGATE THE LEXICAL AND GRAMMATICAL ENVIRONMENT OF COGNATES ACROSS ENGLISH AND ITALIAN

1. Introduction

Word Sketch, “an automatic, corpus-derived summary of a word’s grammatical and collocational behaviour” (Kilgarriff et al. 2010) and part of the battery of search strategies provided by the Sketch Engine¹ (Kilgarriff et al. 2014), was essentially a reaction to a number of corpus query requirements. Concordances were a pioneering invention but they can prove unwieldy, difficult to read and thus time-consuming (Kilgarriff and Kozem 9-13; Atkins and Rundell 109), even with alphabetical sorting. Collocational profiles are also useful for investigating the immediate environment of a word, but tend to be in the form of a simple list based upon a relationship of frequency or salience between the headword and co-occurrences within an arbitrary window of text and with no distinction of the grammatical relations at work in the sentence. In any case this type of collocation finding “is grammatically blind. It only considers proximity” (Kilgarriff and Kozem 14, see also Thomas 2015). This of course may be sufficient for given users, but Word Sketch (henceforth WS) at once appealed in particular to lexicographers because for the first time it provided systematic lists of collocates divided according to the grammatical sequence in which they appear:

The word sketch [...] provides one list of collocates for each grammatical relation the word participates in. For a verb, the subject, the objects, the conjoined verbs (*stand and deliver, hope and pray*), modifying adverbs, prepositions and prepositional objects, are all presented in different lists. (Kilgarriff et al. 2004)

This paper focuses on examples of WS searches across English and Italian, with particular reference to cognates of the two languages with apparently similar meanings, in an attempt to shed light on the advantages and disadvantages of WS to investigate cross-language near-synonym differentiation, primarily with language learners and translators in mind. To familiarise ourselves with WS queries let us consider a couple of fairly straightforward examples.

2. Word Sketch: preliminary examples: *off-centre* and *foresee*

The first example is the adjective *off-centre* in the enTenTen corpus.² In its hyphenated form (I shall discuss the question of hyphenated and unhyphenated forms later in this paper) it occurs 699 times in WS with a relatively low frequency ratio of 0.05 occurrences per million words in the corpus, within a very limited range of grammatical relations (Fig.1).

* *Dominic Stewart teaches English Language and Italian-English Translation at the University of Trento. His research interests include corpus linguistics and translation into a foreign language. He is the author of Semantic Prosody: a critical evaluation (2010) and Translating Tourist Texts from Italian to English as a Foreign Language (2012).*

¹ The Sketch Engine is a corpus manager and analysis software created by Lexical Computing Ltd in 2003, now with over 300 corpora in 80 languages. See (<https://www.sketchengine.co.uk>) for further details.

² In this paper I use two corpora from the TenTen corpus family (see Jakubiček et al. 2013) which comprises comparable web-based large corpora, including those used in the present analysis, i.e., enTenTen, over 11 billion words of English collected in 2012, and itTenTen, 2.5 billion words of Italian collected in 2010. There is a considerable difference in the size of the two corpora, and I therefore always report the respective frequency ratios when comparing them. The 2013 version of enTenTen was avoided because it is so massive (nearly 20 billion words) that it would have dwarfed the itTenTen and made the data less comparable as a result.



Fig.1: Word Sketch for *off-centre* in enTenTen
off-centre (adjective)
enTenTen [2012] freq = 699 (0.05 per million)

adj_subject		
	98	2.10
Arch	7	1.36

modifier		
	116	0.10
excellently	2	2.72
curiously	2	2.31
slightly	55	1.71
noticeably	2	1.09

modifies		
	351	0.30
indierock	3	7.37
noisepop	2	7.13
indiepop	5	7.05
Lacing	2	3.03
trackpad	2	2.75
turning	2	2.09
Hit	53	1.11
Fin	3	0.10

and/or		
	106	0.20
caudal	2	4.21
Jazzy	2	2.48
Quirky	2	0.19

Noticeable collocations are *slightly off-centre*, where *slightly* is a “modifier” of *off-centre*, and in the “modifies” column *off-centre hit* (WS captures the lemma rather than the single word form so *hit* in reality includes both *hit* and *hits* – see 7.2 below for discussion). Clicking on these will take us to the attendant concordances, where we note that the first sequence is predominantly associated with music, and that the second (if we click on “text types”) is almost always connected with golf and appears in golfing magazines. Also worthy of note is the co-occurrence with *arch* within the grammatical relations (gramrels) column “adj_subject”, for example *the broad chancel arch is off-centre and relatively low*, all 7 occurrences here belong to the language of architecture. Finally, the grammatical relation “and/or” captures words which combine with the headword by means of *and/or* or within a list/sequence, e.g., *an engagingly lavish set replete with quirky, off-centre atmospheres*. This column can be useful to identify semi-synonymous words (*kind and generous*), though of course the items in the column can bear a completely opposite meaning (*right or wrong*).

With this preliminary, relatively low-frequency example I have deliberately avoided a complex WS, but it can already be understood how this layout of information – subdivided as it is into collocations through grammatical relations – might in many circumstances be more digestible than a list of concordance lines. The sketch in question seems useful not only for lexicographers but also for students of English (see Carloni 2015) or for translators into English as a foreign language, who might as a result be dissuaded from adopting *off-centre position* or *off-centre location* to render the collocation *posizione decentrata* in a sequence such as “cadde in disuso per la sua posizione decentrata rispetto alle zone abitate della città.”



As a further preliminary example, let us examine a WS for the verb *foresee* (Fig.2), which retrieves over 51,000 occurrences with a frequency ratio of 3.97 hits per million words in the corpus (the underlined words are clickable and enable you to access a separate WS of the keyword + the collocate in question).

Fig.2: Word Sketch for *foresee* in enTenTen

foresee (verb)

enTenTen [2012] freq = [51,509](#) (3.97 per million)

<u>object</u>			<u>subject</u>			<u>modifier</u>			<u>and/or</u>			<u>ing_comp</u>		
	27,062	0.70		10,761	0.40		10,067	0.10		1,502	0.00		1,377	0.40
Eventuality	43	4.79	Raiden	13	5.01	Sensibly	33	5.54	Foreknow	9	6.56	pertain	16	0.15
<u>future</u>	1,010	3.74	Nostradamus	15	4.49	<u>reasonably</u>	256	5.19	Foreordain	9	6.51	<u>pro_object</u>		
Doom	36	3.65	Framers	8	4.30	Prophetically	8	4.43	Foretell	56	6.08	1,350 0.30		
<u>consequence</u>	352	3.47	Huxley	15	4.24	Misleadingly	7	4.23	Forestall	34	5.81	<u>myself</u>	120	1.97
happening	45	3.28	Kurzweil	8	3.98	Infallibly	7	4.22	Avert	8	2.08	yourself	74	0.53
<u>danger</u>	208	3.10	prognosticator	7	3.97	Dimly	13	4.16	Predict	55	1.93	<u>pp_in-i</u>		
single-driver	7	3.07	Trotsky	14	3.88	Accurately	77	4.08	Anticipate	33	1.62	1,217 0.20		
<u>possibility</u>	346	3.00	Isaiah	32	3.87	Wisely	16	3.29	<u>prevent</u>	112	1.17	Article	15	1.20
collapse	79	2.98	Ezekiel	16	3.84	Realistically	9	3.02	Guard	9	0.56	Treaty	7	0.79
catastrophe	33	2.95	Visionary	19	3.65	Correctly	67	2.98	<u>pro_subject</u>			directive	7	0.47
calamity	19	2.91	Prophet	91	3.53	Falsely	11	2.93	9,711 0.50			future	89	0.25
contingency	23	2.78	Engels	9	3.47	Incorrectly	10	2.61	<u>he</u>	1,941	1.12	Model	7	0.02
slowdown	21	2.77	Marx	30	3.41	Neither	21	2.59	<u>!</u>	2,773	0.27	<u>pp_by-i</u>		
demise	32	2.77	Orwell	11	3.32	Plainly	10	2.48	<u>we</u>	1,212	0.16	961 0.60		
inevitability	9	2.75	Nietzsche	9	2.97	<u>possibly</u>	128	2.30	<u>wh_comp</u>			Marx	8	1.67
wickedness	11	2.68	Economist	70	2.96	<u>clearly</u>	127	2.23	2,142 1.70			prophet	14	0.88
pitfall	21	2.61	Astrologer	9	2.95	Expressly	9	2.21	<u>which</u>	153	2.22	Article	8	0.29
destiny	36	2.51	Forecaster	8	2.86	Precisely	27	2.10	<u>what</u>	968	2.15	<u>pp_at-i</u>		
layoff	16	2.35	Pundit	17	2.70	Rightly	9	1.95	Whenever	14	2.12	208 0.10		
evil	46	2.31	no-one	13	2.38	Initially	37	1.83	<u>how</u>	545	1.78	outset	9	2.21
shortage	57	2.30	Analyst	95	2.29	Ago	20	1.69	<u>when</u>	269	1.21	<u>pp_with-i</u>		
<u>outcome</u>	186	2.25	Nobody	72	2.08	Explicitly	10	1.55	Where	93	1.11	170 0.10		
advent	19	2.23	Fathers	8	1.76	Likewise	14	1.38	That	40	0.67	certainty	25	2.52
cutback	9	2.23	Poe	7	1.64	Exactly	74	1.38	Who	33	0.66			
resurgence	10	2.19	Planner	27	1.48	<u>not</u>	3,496	1.33						

<u>pp_from-i</u>		
	139	0.10
Eternity	10	2.08

<u>pp_if-i</u>		
	59	0.40
metaphysicalconsiderations	17	11.97

In this case the grammatical relation columns are more prolific, the richest of these being “object” (*foresee the possible dangers*), “subject” (*the prophets foresaw*), “modifier” (*clearly foresee*), “and/or” (*foreseen and prevented*), while others include “pp_in-i” (*foresee in the future*) and “wh_comp” (*foreseen which*). I shall focus on the first two. The first column lists salient grammatical objects, and the second column lists salient grammatical subjects. What is immediately obvious is that the subject of *foresee* is typically a person, while



the object is often an unfavourable scenario. This distribution is very different from the Italian dictionary equivalent *prevedere*, whose grammatical subject is frequently either a person or a thing, and whose grammatical objects include the duration of an activity, as in the following example from a tourist brochure: “Il primo percorso, lungo poco più di 4 chilometri, di media difficoltà e che prevede un tempo di percorrenza di circa 3 ore, parte da Piazza del Popolo.” This of course is precious information for the non-native speaker of English.

3. Word Sketch beyond the dictionary

Since all modern dictionaries are corpus-based, and since many of them adopt WS, it seems legitimate to wonder whether language operators other than lexicographers need to use WS at all. After all, isn't the information available already incorporated in dictionary entries? The answer to this question is yes, but to a limited degree. Dictionaries are by nature concise, so lexicographers are required to condense a vast quantity of information into a very restricted window. Clearly the more typical structures and collocations are prioritised, but these do not necessarily rule out other usage. For example the *Macmillan Dictionary Online* defines *foresee* as “to see or know something that will happen in the future” and supplies just one example: “Who could have foreseen such problems?”, but this terse definition (perhaps unsatisfactory anyway because it rests on the dubious notion that we can know future events) with just a single example is in no way sufficient to suggest exclusion of a sequence such as “the itinerary foresees a duration of 3 hours” – a literal and poor translation of the Italian sentence in Section 2 above – particularly if the itinerary is a proposal rather than an established route.

Further, dictionaries may struggle to capture possible pragmatic associations of a given word or expression. The *Macmillan Dictionary Online* provides pragmatic information about the adverb *utterly*: “completely: often used for emphasising how bad someone or something is: *You're being utterly unreasonable,*” and this is reflected in the unfavourable meaning of this word's most typical co-occurrences (for instance *utterly ridiculous / pointless / useless / worthless*). Yet the *Macmillan Dictionary Online* (the first dictionary ever to adopt Word Sketch, see Kilgarriff and Rundell) does not report the possible pragmatic associations of another adverb with a similarly unfavourable immediate lexical environment, that is *singularly*, whose definition is as follows:

singularly

in a noticeable way

The committee is singularly impressed at the originality of your research.

A WS for this adverb (9,756 hits, 0.8 per million) suggests that it typically “modifies” adjectives describing unpleasant scenarios, for example *unimpressed, unhelpful, inept, unattractive, unsuccessful* and *obtuse*, while barely any pleasant-sounding words appear in this column. Yet the *Macmillan* supplies no pragmatic information about this word, listing just one example (“singularly impressed”) which is in any case unrepresentative of this adverb's immediate lexical environment.

The same of course goes for Italian dictionaries too. For example the *Sabatini Coletti Online* includes the following entry for another adverb, *bellamente*:

avv. bellamente 1. Garbatamente, gentilmente 2. Pacificamente, con calma: *se ne andava bellamente per la sua strada; a buon diritto: quest'opera può bellamente figurare tra i capolavori del secolo.*

A WS deriving from the itTenTen corpus (2270 occurrences of *bellamente*, 0.73 per million) shows that it typically modifies verbs rather than adjectives. These verbs are predominantly characterised by unfavourable meanings: *infischiare/infischiarsi, impippa, fregare/fregarsene, fottere/fottersene, snobbare* while others include *ignorare, aggirare, tralasciare* and *calpestare*. Of course not all of these verbs have inherently unfavourable meanings (for example *ignorare* can have a fairly innocent meaning of “not know / not be



informed”), so it is advisable to check the respective concordances for further context, but the habitually unfavourable environment of *bellamente* – barely suggested by the dictionary – is hard to dispute.

4. Semantic prosody

Meaning which is said to be read off from the semantic preference(s), or at least from the immediate lexical environment, of a word or expression has primarily been known as semantic prosody, a concept which falls squarely within the post-Firthian corpus linguistics ethos of using collocational information to infer word sense within discourse. Semantic prosody is almost always polarised by scholars in terms of positive/favourable vs negative/unfavourable. According to this dichotomy, *bellamente* would be assigned an unfavourable prosody in view of its unfavourable-sounding semantic preference of “not giving a damn,” and *singularly* would be also assigned an unfavourable prosody in view of its habitually unpleasant immediate lexical environment. These unfavourable associations are then regarded as part of the meaning of *singularly* and *bellamente*, derived from their habitual company. This hypothesis of transfer of meaning from lexical company to the node is controversial; some scholars, notably Whitsitt, affirm that it will never be possible to prove there has been a transfer of meaning from one word to another, it is simply that corpus data tend to make us see things which are not present. Further, if *utterly*, *singularly* etc. do have this transferred negative meaning, why is this not brought to bear in, for example, the combinations *singularly beautiful* or *utterly compelling*?

The concept of semantic prosody has further theoretical complications, and scholars cannot agree on how to define or approach it (see Stewart 2010 for a critique of descriptions of semantic prosody over the last 25 years). For example, while the arguments above concerning *bellamente* and *singularly* appear to take it as read that semantic prosody characterises the word, other interpretations focus on its pragmatic function and how it expresses the attitude of the speaker or writer towards a pragmatic situation across longer stretches of discourse. Even the notion of “semantic preference” is controversial (Stewart 2010, 89-91), but since my principal concern in this paper is to engage with the functionings of WS it is not my intention to tackle these various complications here. Suffice it to say that – in accordance with the structure of WS – I shall focus on the immediate lexical and grammatical environment of words and expressions as a means to better understand in what co-texts and contexts the keyword is habitually used.

5. Near-synonyms across languages

Over the last 15 years or so a number of studies have been devoted to the investigation of near-synonyms between English and other languages by means of corpus analysis, e.g., Berber-Sardinha for Portuguese; Dam-Jensen and Zethsen for Danish; Munday for Spanish; Stewart (2009), Olohan 35-39 for Italian; McEnery and Xiao, Xiao and McEnery, Lee and Liu for Chinese. Inevitably, findings have for the most part demonstrated that apparent synonyms, whether within or across languages, have distinct colligational and collocational profiles, re-emphasising the view that perfect synonymy does not exist. These studies, however, are conducted mostly with the use of concordances or simple lists of collocates. My focus in this paper is not to re-raise theoretical issues of imperfect synonymy but to go a step further in terms of corpus analysis, verifying to what degree WS facilitates the study of the lexical and grammatical environment of cognates across languages, whether for language-learning or translational purposes.

6. Cognates across Italian and English

In the present paper I shall focus on words in English and Italian which would appear to be almost mirror images of one another on both a formal and a semantic level, i.e., physically similar and semantically similar, of the type *mysterious / misterioso*. I thus exclude (i) physically similar pairings such as *actual / attuale*, *gymnasium / ginnasio* because despite their formal resemblance they have very dissimilar meanings (and are thus frequently denominated “false friends”), (ii) semantically similar pairings such as *sad / triste*, because notwithstanding their analogous meanings they are formally very distinct, and (iii) pairings such as *mouse / topo*, because despite a basic semantic similarity one of them has developed a highly specific meaning in a specific sector (*mouse* in the language of computers). This method of comparing close cognates, it should be stressed, is in part no more than a convenience designed to highlight the pros and



<u>unashamedly</u>	<u>6</u>	5.26	<u>spa</u>	<u>86</u>	2.95	<u>atteggiamento</u>	<u>27</u>	1.96	<u>pp_verso-i</u>		
<u>wickedly</u>	<u>6</u>	4.73	<u>pudding</u>	<u>17</u>	2.76	<u>simpatia</u>	<u>6</u>	1.85		<u>151</u>	<u>154.90</u>
<u>luxuriously</u>	<u>6</u>	4.64	<u>pedicure</u>	<u>6</u>	2.67	<u>sguardo</u>	<u>24</u>	1.72	<u>debolezza</u>	<u>6</u>	1.71
<u>gloriously</u>	<u>7</u>	4.59	<u>smile</u>	<u>128</u>	2.58	<u>saggezza</u>	<u>3</u>	1.24	<u>pp_nel-x</u>		
<u>wonderfully</u>	<u>43</u>	4.28	<u>brunch</u>	<u>9</u>	2.37	<u>papà</u>	<u>5</u>	1.21		<u>131</u>	<u>11.50</u>
<u>fabulously</u>	<u>6</u>	4.27	<u>escape</u>	<u>34</u>	2.20	<u>verso</u>	<u>10</u>	1.07	<u>confronto</u>	<u>118</u>	2.45
<u>overly</u>	<u>67</u>	4.23	<u>getaway</u>	<u>27</u>	2.18	<u>comprensione</u>	<u>7</u>	0.99	<u>pp_del-x</u>		
<u>ridiculously</u>	<u>12</u>	3.38	<u>concoction</u>	<u>6</u>	1.93	<u>tolleranza</u>	<u>3</u>	0.86		<u>32</u>	<u>0.60</u>
<u>delightfully</u>	<u>5</u>	3.30	<u>brownie</u>	<u>6</u>	1.85	<u>ironia</u>	<u>4</u>	0.77	<u>tribunale</u>	<u>5</u>	0.49
<u>richly</u>	<u>7</u>	2.81	<u>bath tub</u>	<u>9</u>	1.83	<u>occhio</u>	<u>23</u>	0.16			
<u>excessively</u>	<u>7</u>	2.48	<u>massage</u>	<u>56</u>	1.75						
<u>purely</u>	<u>21</u>	2.37	<u>velvet</u>	<u>7</u>	1.75						
<u>suitably</u>	<u>5</u>	2.27	<u>spree</u>	<u>7</u>	1.74						
<u>massively</u>	<u>5</u>	2.17	<u>Parent</u>	<u>6</u>	1.74						
<u>utterly</u>	<u>16</u>	2.09	<u>Spa</u>	<u>16</u>	1.73						
<u>downright</u>	<u>5</u>	2.06	<u>grin</u>	<u>10</u>	1.56						
<u>terribly</u>	<u>8</u>	1.63	<u>chocolate</u>	<u>57</u>	1.49						
<u>over</u>	<u>27</u>	1.47	<u>extra</u>	<u>10</u>	1.49						
<u>uniquely</u>	<u>6</u>	1.12	<u>amusement</u>	<u>10</u>	1.48						
<u>similarly</u>	<u>7</u>	1.06	<u>nostalgia</u>	<u>5</u>	1.47						

<u>and/or</u>	<u>3,419</u>	<u>0.30</u>	<u>e_o</u>	<u>605</u>	<u>1.80</u>
<u>self-</u>	<u>10</u>	6.10	<u>clemente</u>	<u>13</u>	7.92
<u>permissive</u>	<u>16</u>	5.37	<u>salvatrice</u>	<u>4</u>	7.24
<u>decadent</u>	<u>24</u>	4.88	<u>autoindulgente</u>	<u>3</u>	7.17
<u>neglectful</u>	<u>7</u>	4.83	<u>tollerante</u>	<u>16</u>	6.58
<u>hedonistic</u>	<u>6</u>	4.43	<u>permissivo</u>	<u>7</u>	6.40
<u>amused</u>	<u>7</u>	4.35	<u>accomodante</u>	<u>3</u>	5.71
<u>opulent</u>	<u>13</u>	4.34	<u>benevolo</u>	<u>11</u>	5.67
<u>undisciplined</u>	<u>5</u>	4.10	<u>bonario</u>	<u>7</u>	5.51
<u>luxurious</u>	<u>85</u>	3.92	<u>comprensivo</u>	<u>24</u>	5.21
<u>relaxing</u>	<u>40</u>	3.81	<u>misericordioso</u>	<u>4</u>	5.14
<u>pretentious</u>	<u>9</u>	3.73	<u>compassionevole</u>	<u>3</u>	4.82
<u>authoritarian</u>	<u>16</u>	3.73	<u>divertito</u>	<u>6</u>	4.81
<u>buttery</u>	<u>7</u>	3.68	<u>premuroso</u>	<u>4</u>	4.66
<u>sumptuous</u>	<u>10</u>	3.46	<u>caritatevole</u>	<u>3</u>	4.63
<u>extravagant</u>	<u>16</u>	3.37	<u>intransigente</u>	<u>5</u>	4.53
<u>self-serving</u>	<u>7</u>	3.34	<u>mite</u>	<u>11</u>	4.32
<u>creamy</u>	<u>25</u>	3.31	<u>tantino</u>	<u>3</u>	4.05
<u>narcissistic</u>	<u>5</u>	3.27	<u>affettuoso</u>	<u>10</u>	3.76
<u>selfish</u>	<u>25</u>	3.05	<u>severo</u>	<u>17</u>	3.69
<u>home-made</u>	<u>5</u>	3.03	<u>comosso</u>	<u>3</u>	3.62
<u>lavish</u>	<u>12</u>	2.92	<u>protettivo</u>	<u>8</u>	3.59
<u>tolerant</u>	<u>15</u>	2.90	<u>amorevole</u>	<u>3</u>	3.57
<u>wasteful</u>	<u>8</u>	2.88	<u>pietoso</u>	<u>4</u>	3.55
<u>scrumptious</u>	<u>5</u>	2.86	<u>generoso</u>	<u>9</u>	2.99
<u>affectionate</u>	<u>10</u>	2.82	<u>spietato</u>	<u>4</u>	2.53

What strikes the user at once is the imbalance of the resulting sketch, notwithstanding the similar frequency ratios: *indulgent* (12,224 hits, 0.94 per million) is possessed of a modifier column (*deliciously*, *overly*,

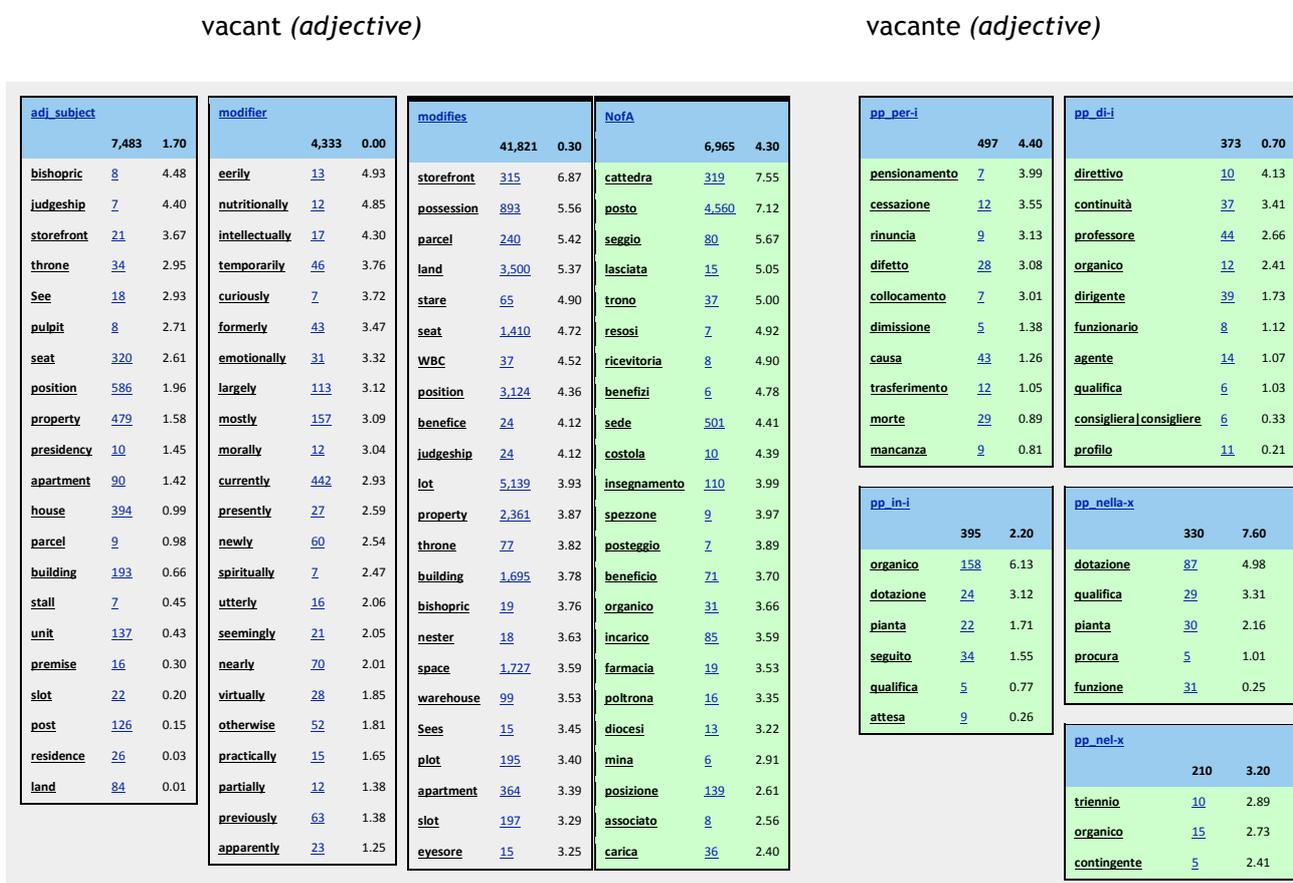


wonderfully) whereas *indulgente* (2,748 hits, 0.89 per million) is not. We do, however, find columns of nouns modified by the two adjectives, where a contrast is already manifest (for *indulgent* the list includes a host of tangible things such as *pudding, dessert, spa, massage, chocolate*, while for *indulgente* we find more abstract notions like *tenerezza, atteggiamento, comprensione*), but the most striking divergence is to be noted in the “and/or” columns, in which the Italian adjectives listed – with just one or two exceptions – have a tender feel to them (*clemente, comprensivo, benevolo, tollerante, mite*), whereas the English list has a far higher percentage of adjectives representing undesirable qualities (*selfish, authoritarian, pretentious, self-serving, wasteful*) though there are certainly favourable elements too (*tolerant, relaxing, affectionate*). Naturally the “and/or” column does not feature only semi-synonymous terms – as stated earlier it frequently features opposites (*thick and thin, right or wrong*) – but it can disclose important information all the same. The reason for the substantial difference between the respective Word Sketches is that *indulgent* so often means *self-indulgent, self-pampering*, whereas *indulgente* simply has the meaning of *lenient* or *sympathetic*. However, despite the fact that this Bilingual Word Sketch throws up some surprises, it still does not go completely beyond the dictionary. Although for *indulgent* the *Macmillan Dictionary Online* provides only the meaning “allowing someone to do or have what they want,” the *Oxford Advanced Learner’s Dictionary Online* supplies the indication “See also self-indulgent,” and the second meaning of *indulgent* listed by the *Oxford Dictionary Online* is “self-indulgent.”

6.2.2 vacant vs vacante

In terms of the nouns they modify the most obvious difference between these two adjectives is that *vacant* (68,161 hits, 5.25 per million) combines much more often with buildings than *vacante* (11,055 hits, 3.59 per million, and thus less frequent), which typically modifies a position, be it professional, political or hereditary: *incarico, cattedra, insegnamento, seggio, trono, sede* (Fig.4).

Fig.4: Bilingual Word Sketch for *vacant / vacante* in enTenTen / itTenTen





6.2.3 lucidity vs lucidità

For this final example I would like to begin with the respective definitions in the *Oxford Dictionary Online* and the *Sabatini Coletti Online*:

lucidity

1 Clarity of expression; intelligibility *His lecture combined intellectual lucidity and passion*

1.1 The ability to think clearly, especially in intervals between periods of confusion or insanity
She had moments of lucidity

lucidità

visione intellettuale chiara e distinta, perfetta consapevolezza di sé e delle cose: *affrontare un problema con lucidità* || lucidità di mente, pieno possesso delle proprie facoltà mentali

The respective definitions lie along parallel semantic lines apart from the allusion in the *Oxford* to “intervals between periods of confusion or insanity” – which I shall return to below – and the same is true of the definitions of *lucid* and *lucido*. Nevertheless, the Bilingual Word Sketch *lucidity* (4,695 hits, 0.36 per million) vs *lucidità* (14,590 hits, 4.74 per million and thus proportionally far more frequent than its English counterpart) would suggest that the two words are not as parallel as they seem (Fig.5).

Fig.5: Bilingual Word Sketch for *lucidity* / *lucidità* in enTenTen / itTenTen

lucidity (noun)

enTenTen [2012] freq = [4,695](#) (0.36 per million)

lucidità (noun) itTenTen [2010] freq = [14,590](#) (4.74 per million)

Use another candidate

translation: [sprazzo](#) [comizio](#) [sorpassare](#) [stupido](#) [Damasco](#) [lucido](#) [Paolo](#) [disarmare](#) [Schmidt](#)

Click on collocates to access reciprocal bilingual search or find [translated collocations](#)

object of		preN V		subject of		postN V	
	987 0.20		2,979 2.10		420 0.10		1,031 1.20
concede	9 1.96	riacquistare	43 6.56	widen	4 0.38	sottoporta	6 7.20
commend	4 0.98	offuscare	15 5.47			disarmare	21 6.11
regain	10 0.98	appannare	8 5.46			analizzare	8 0.70
attain	20 0.76	smarrire	18 5.43			decidere	9 0.18
induce	9 0.03	annebbiare	8 5.37			dimostrare	11 0.06
		disarmare	11 4.92			affermare	6 0.05
		perdere	505 4.85				
		perdere perdonare	31 4.63				
		mantenere	207 4.38				
		impressionare	7 4.17				
		riconquistare	7 4.00				
		togliere	88 3.83				
		denotare	8 3.82				
		recuperare	57 3.71				
		ritrovare	70 3.58				



<u>conservare</u>	<u>65</u>	3.53
<u>compromettere</u>	<u>15</u>	3.48
<u>ammirare</u>	<u>19</u>	3.24
<u>alterare</u>	<u>9</u>	3.02
<u>apprezzare</u>	<u>29</u>	2.78
<u>accentuare</u>	<u>5</u>	2.66
<u>ridare</u>	<u>6</u>	2.63
<u>mancare</u>	<u>77</u>	2.63
<u>donare</u>	<u>12</u>	2.52
<u>dimostrare</u>	<u>60</u>	2.50

<u>modifier</u>		
	2,039	0.20
<u>initiated</u>	<u>4</u>	5.73
<u>phenomenal</u>	<u>22</u>	3.14
<u>admirable</u>	<u>8</u>	2.37
<u>unparalleled</u>	<u>10</u>	2.27
<u>low-level</u>	<u>4</u>	2.19
<u>utmost</u>	<u>10</u>	2.05
<u>coloration</u>	<u>8</u>	1.75
<u>startling</u>	<u>5</u>	1.73
<u>heightened</u>	<u>4</u>	1.70
<u>terrific</u>	<u>18</u>	1.54
<u>customary</u>	<u>7</u>	1.53
<u>sonic</u>	<u>5</u>	1.51
<u>high-level</u>	<u>4</u>	1.08
<u>remarkable</u>	<u>21</u>	0.91
<u>singular</u>	<u>4</u>	0.77
<u>spontaneous</u>	<u>4</u>	0.76
<u>astonishing</u>	<u>4</u>	0.69
<u>superb</u>	<u>16</u>	0.69
<u>optical</u>	<u>7</u>	0.65
<u>extraordinary</u>	<u>14</u>	0.52
<u>clarity</u>	<u>12</u>	0.46
<u>tremendous</u>	<u>12</u>	0.41
<u>aesthetic</u>	<u>5</u>	0.41
<u>mental</u>	<u>32</u>	0.39
<u>incredible</u>	<u>20</u>	0.32

<u>n_modifier</u>		
	137	1.00
<u>analisi</u>	<u>19</u>	0.12

<u>modifies</u>		
	401	0.00
<u>measuremen</u>	<u>5</u>	6.42
<u>rrnside</u>	<u>8</u>	5.51

<u>and/or</u>		
	1,530	0.30
<u>caret</u>	<u>11</u>	6.53
<u>measuremen</u>	<u>5</u>	5.75
<u>conciseness</u>	<u>6</u>	5.61
<u>directness</u>	<u>8</u>	4.96
<u>profundity</u>	<u>6</u>	4.75
<u>clearness</u>	<u>6</u>	3.96
<u>brevity</u>	<u>5</u>	3.21
<u>clarity</u>	<u>70</u>	3.01
<u>eloquence</u>	<u>5</u>	2.89
<u>carat</u>	<u>9</u>	2.72
<u>detachment</u>	<u>11</u>	2.63
<u>sharpness</u>	<u>6</u>	2.55
<u>coloration</u>	<u>13</u>	2.46
<u>sobriety</u>	<u>4</u>	2.04
<u>alertness</u>	<u>4</u>	1.93
<u>purity</u>	<u>10</u>	1.32
<u>madness</u>	<u>7</u>	1.17
<u>brilliance</u>	<u>5</u>	1.08
<u>precision</u>	<u>12</u>	0.90
<u>simplicity</u>	<u>11</u>	0.80
<u>wit</u>	<u>12</u>	0.71
<u>coloring</u>	<u>6</u>	0.56
<u>openness</u>	<u>4</u>	0.49
<u>restraint</u>	<u>5</u>	0.26
<u>courage</u>	<u>10</u>	0.15

<u>e_o</u>		
	5,278	1.90
<u>freddezza</u>	<u>89</u>	7.54
<u>prontezza</u>	<u>48</u>	7.01
<u>lungimiranza</u>	<u>34</u>	6.31
<u>schiettezza</u>	<u>18</u>	6.06
<u>pacatezza</u>	<u>17</u>	6.01
<u>brillantezza</u>	<u>17</u>	5.88
<u>fermezza</u>	<u>40</u>	5.88
<u>acutezza</u>	<u>15</u>	5.67
<u>obiettività</u>	<u>23</u>	5.64
<u>chiarezza</u>	<u>163</u>	5.60
<u>calma</u>	<u>78</u>	5.58
<u>serenità</u>	<u>84</u>	5.54
<u>onestà</u>	<u>58</u>	5.27
<u>grinta</u>	<u>26</u>	5.25
<u>coraggio</u>	<u>174</u>	5.09
<u>franchezza</u>	<u>12</u>	5.04
<u>freschezza</u>	<u>29</u>	5.02
<u>razionalità</u>	<u>31</u>	4.96
<u>concretezza</u>	<u>26</u>	4.96
<u>concisione</u>	<u>6</u>	4.92
<u>acume</u>	<u>7</u>	4.92
<u>autocontrollo</u>	<u>11</u>	4.85
<u>disincanto</u>	<u>9</u>	4.84
<u>sinteticità</u>	<u>5</u>	4.75
<u>spietatezza</u>	<u>5</u>	4.64

<u>pp_obj_of-i</u>		
	677	0.50
<u>moment</u>	<u>152</u>	0.99
<u>instant</u>	<u>5</u>	0.53



pp_of-i		
	377	0.30
exposition	6	1.64
prose	4	0.51

pp_obj_with-i		
	294	0.90
expound	4	2.32

predicate_of		
	47	0.20
acronym	4	0.89

Particularly eye-catching are the columns capturing adjectives which typically qualify the key words. The Italian list (“AofN”) is dotted with powerful, unpleasant-sounding adjectives, for instance, *spietato*, *implacabile*, *impietoso*, *scarso* and *disincantato*, combinations which habitually describe the ability to react in some way to a sometimes very unpleasant scenario. Noteworthy too is the presence of other adjectives with forceful meaning: *estremo*, *impressionante*, *folgorante*, *sconvolgente*. The English list (“modifier”), on the other hand, seems very upbeat, characterised primarily by positive-sounding collocates such as *phenomenal*, *admirable*, *unparalleled*, *terrific*, *remarkable*, *astounding*, *extraordinary*, *tremendous*, *incredible* and *superb*.

This difference is not in evidence in the “and/or” column, both of which are characterised prevalently by nouns describing favourable qualities such as *clarity*, *eloquence*, *wit / chiarezza*, *coraggio*, *serenità*. Perhaps the only word in the “and/or” column which really stands out is *freddezza* in the Italian list, one of the most frequent collocates of *lucidità* (89 co-occurrences). By clicking on this collocate we are taken to a concordance containing occurrences of both words within a span of 5 to the left and right: *freddezza e lucidità* 34 hits, *lucidità e freddezza* 24 hits, while the remaining occurrences are almost all part of a list of nouns separated by commas (...*freddezza, lucidità, controllo e misura*). Especially striking is that over 20 of these occurrences are connected with violence and/or murder.

On the basis of this evidence, albeit summarily described here, it would seem that *lucidità* is often associated with (sometimes extremely) unpleasant scenarios, though there is no trace of this in dictionaries. *Lucidity*, on the other hand, is not associated with such scenarios, though it often occurs in contexts that entail previous or subsequent periods of mental confusion, a state of affairs included in the *Oxford Dictionary Online* definition, though as far as I can make out, the only hint we have of this state of affairs in WS is its frequency of co-occurrence with *moment* (152 times), often preceded by *brief*, *few* and *rare*. It would thus seem to be the case that the lexical environments of *lucidity* and *lucidità* are very different, despite their similar basic meaning.

7. Drawbacks of Bilingual Word Sketches

The queries carried out so far illustrate the importance of WS in going beyond dictionary definitions, revealing co-texts and nuances not recorded in lexicographical entries and perhaps not imagined by advanced language users. There are however certain methodological drawbacks to comparing different languages through WS, which will be outlined in this section.

7.1 The imbalance of bilingual Word Sketches

Bilingual WS is clearly a precious resource, but the two compared languages will inevitably have different architectures, a factor which may hamper a direct comparison of lexical environment. The juxtaposition, for



example, of the WS of *vacant* and *vacante* turns up some fascinating insights and some minute detail, but there is a substantial imbalance which cannot be attributed simply to the different respective frequency ratios in the two corpora:

- firstly, the nouns accompanying *vacant* are divided into two columns: “adj_subject” (*the premises were vacant*) and “modifies” (*vacant plots of land*), whereas the nouns accompanying *vacante* are assigned just one list, namely “NofA”, which basically corresponds to the English “modifier” column (*sede vacante, posto vacante*), with no evidence of sequences such as *la sede è vacante, i posti sono vacanti* etc.
- secondly, the “modifier” column is present for English (*currently vacant*) but not for Italian (despite the abundance of candidates: *ancora vacante, tuttora vacante, ora vacante, ormai vacante, già vacante*), something which may be due to the fact that *ancora, tuttora, ora* etc. are such high-frequency words that their combination with *vacante* is not sufficiently salient
- thirdly, the Italian part of the WS is dominated by a plethora of columns reporting the keyword alongside prepositional phrases (*i posti vacanti nella qualifica di dirigente*), whereas these are absent on the English side of the WS despite several possible candidates in the enTenTen corpus (for instance a simple query for *vacant positions in* returns 155 occurrences)
- lastly, the comparison of the two “and/or” columns for *vacant* and *vacante* is skewed by the presence of a series of adjectives in the Italian list which are in reality not “and/or” nor part of a list of adjectives but are instead simply an item of a common collocation: the presence of *dirigenziale* stems from the collocation *posto dirigenziale* (hence sequences such as *posto dirigenziale vacante*), *farmaceutico* from the collocation *sedi farmaceutiche*, *ordinario* from *professore ordinario* etc. Thus *dirigenziale, farmaceutico* and *ordinario* have no real entitlement to appear in the “and/or” column of *vacante*. The English “and/or” list is more authentic, so to speak, containing sequences such as *the clearance of vacant or derelict land, to purchase the vacant, run-down house next to her, a previously vacant and neglected building*, though there are a few examples of the modifier type, for example *vacant managerial post*. It should be underlined that these differences apply even when two separate WS queries are performed, one for *vacant* and one for *vacante*.

7.2 Lemmatisation

WS queries are enabled only for lemmas and not for single word forms (for example the lemma *confirm* embraces the word forms *confirm, confirms, confirming* and *confirmed*), thus WS comparisons hinge upon whether words have been (automatically) tagged as lemmas or not, but lemmatisation is not always consistent. A bilingual WS for the adjectives *sconcertante / disconcerting* is skewed by the fact that while the adjective *sconcertante* is tagged as a lemma in itTenTen, the adjective *disconcerting* is not tagged as a lemma in enTenTen. Therefore, since WS queries – unlike concordance queries – handle only lemmas, the result is that *disconcerting* can be retrieved solely under the lemma *disconcert*, which effectively means that the only WS comparison we can perform here is between the lemma *sconcertante* (adj) on the one hand and the lemma *disconcert* on the other. The difficulty is of course that the lemma *disconcert* will retrieve not only the form *disconcert* but also *disconcerting, disconcerts* and *disconcerted*, something which prevents a meaningful comparison.

Users need to be aware of this question of lemmatisation since otherwise the results obtained may be hard to understand. For example, while a direct comparison between *disconcerting* and *sconcertante* is not possible, a direct comparison is instead possible for *disconcerted* vs *sconcertato*, since both these words are lemmatised in the respective corpora.

7.3 Contents of the grammatical relations columns

The gramrel columns may feature disparate grammatical structures. The brief “postN_V” (verbs following the key word) column of WS *lucidità* lists five verbs: *disarmare, dimostrare, decidere, analizzare* and *affermare* (as well as *sottoporta*, erroneously tagged as a verb). However, a closer reading reveals that *disarmare* is all adjectival in the form of the present participle (*lucidità disarmante*), *dimostrare* is almost always the past participle (*la scarsa lucidità dimostrata*), *decidere* appears in mostly finite forms (*in un momento di lucidità*



decide di...), *analizzare* is mostly infinitive preceded by *nell'* (*questa lucidità nell'analizzare i fatti*), and *affermare* can occur as a reflexive gerund without a direct connection to the key word (*rivela a scuola lucidità intellettuale affermandosi rapidamente*).

The reason for the absence of *disarmante*, *dimostrato* etc. in the gramrel columns is that the dependence of WS on lemmatisation applies not only to the word queried but also to the sketch itself, i.e., only lemmas can appear in the gramrel columns.

Since neither *disarmante* nor *dimostrato* is lemmatised, they are not eligible for a place in the columns; only the (lemmatised) infinitive form is available. Thus these words, like *disconcerting* above, (i) cannot be the search word in WS, and (ii) cannot figure in any WS columns, whatever the headword may be.

This means that a simple perusal of the gramrels in question can be quite misleading, perhaps deceiving the user into thinking that the words listed are found in similar structures and thus constitute perfectly comparable data.

7.4 Two-word queries

Two-word queries are usually possible with a monolingual WS: separate searches for *black hole*, *tax number*, *buco nero*, *codice fiscale* retrieve results, though queries with more than two words, even high-frequency combinations, produce either very limited results, e.g., *primary school teacher* (even though this sequence has 2635 hits in enTenTen via a simple concordance query), or more often than not no results at all, e.g., *high school senior* (11552 hits via a simple concordance query). However, at the time of writing a bilingual WS is not enabled to compare any word combinations at all, e.g., *black hole* vs *buco nero*, a fact which clearly reduces its range and power.³ Generally speaking WS is weaker when dealing with combinations of words rather than single words, but see Kilgarriff et al 2012.

7.5 Hyphenated words

A bilingual WS can be complicated by the fact that so many lexical items in English have both hyphenated and unhyphenated forms, often with no difference of meaning. A comparison, for example, of the Italian adjective *decentrato* with the English adjective *off-centre* would need to take into account that the unhyphenated *off centre* is also used as an adjective, necessitating a separate WS.

7.6 Drawbacks of Bilingual WS which apply to WS in general

7.6.1 Headings of the gramrel columns

The codes employed for the gramrel columns are sometimes opaque. For instance a WS for *convenient* retrieves a grammatical column with the heading 'np_adj_comp_of', which rather unexpectedly turns out to denote sequences such as *arrange a time convenient for you* and *find it more convenient*.

7.6.2 Initial upper case

WS captures words either with initial upper case or without initial upper case, but not both at the same time. While the simple query (i.e., not the WS query) *codice fiscale* captures both *codice fiscale* and *Codice Fiscale*, a WS for *codice fiscale* captures only initial lower case, and a WS for *Codice Fiscale* retrieves only initial upper case. In the case in point the difference is not critical, but in other cases the discrepancy is enormous. For example, if you are not aware of this upper/lower-case distinction between simple query and WS, then you may be bewildered to discover that a simple query for *christmas* returns almost 900,000 hits, whereas a WS for *christmas* (i.e., with initial lower case) returns only 63,000. The danger is, of course, that

³ In a recent communication (January 2016) the Sketch Engine team informed me that they plan to enable two-word Bilingual WS queries in the near future.



the WS user remains unaware of this and works as a result on only a small percentage of the data available.⁴

7.6.3 Good and bad collocates

One of the problems inherent in assessments of favourable / unfavourable lexical environment in corpora – however one queries the data – is that of whether co-occurrences are to be considered good or bad. While it seems safe to assert that *wonderful* and *splendid* denote good qualities, and that *lousy* and *disgusting* denote bad qualities (though of course in discourse they may be used ironically), it is clear that the interpretation of others depends on contextual factors. It has already been noted that the verb *ignorare* – within the WS of *bellamente* – may have either a neutral or a negative meaning, even though its paradigmatic relationship with verbs such as *infischiare/infischarsi*, *impippa*, *fregare/fregarsene* may persuade us to see it in a negative light. Take also the occurrence of *permissive* within the WS of *indulgent*. The definition of *permissive* supplied in the *Oxford Dictionary Online* – “allowing or characterised by great or excessive freedom of behaviour” – could of course be viewed in both a positive and negative light. The implications of this issue, discussed by Dilts and Newman and by Stewart 2010, 91-97, are clearly not restricted to WS queries, but it helps to be aware of them when evaluating favourable / unfavourable word environment.

7.7. Good use, good results

Using WS is perhaps rather like driving a Ferrari. Its aesthetic appeal may blind you to the fact that you're dealing with highly sophisticated machinery and that you need to know how to handle it. Only good use will produce good results, so it is advisable to be aware of both its pluses and minuses. Having said that, some of the minuses are not sufficiently explained on site, and the Sketch Engine team are currently working to improve the clarity of the interface.

8. Conclusions: WS for learners and translators

Such are some of the advantages and disadvantages of WS in the investigation of cognates across languages. As emphasised above, WS was designed primarily with lexicographers in mind, but how useful is it for learners and translators, or more specifically how useful might the searches performed in this paper be for advanced learners and translators? As regards learners, while it is clear that dictionaries cannot include all the semantic and pragmatic nuances of words and expressions, language learners – even advanced language learners – may take the view that what is absent from a dictionary must be somehow of secondary importance. Bilingual dictionaries account for the semantic and collocational differences of, say, *gravity* vs *gravità* and *sympathy* vs *simpatia*, and such differences will be crucial for learners, but the contrasts between *lucidity* and *lucidità*, beyond the dictionary because more subtle and more nuanced, may not be considered critical by learners trying to stay afloat within that vast ocean of fine distinctions making up a foreign language. For translators, on the other hand, perhaps particularly within the domains of literature and advertising where finding the *mot juste* can be of paramount importance, the benefits of awareness of a word's lexical environment are more readily applicable.

However that may be, it seems important to be aware that precise collocational and colligational mapping between words across languages is a rarity, indeed discrepancies are the norm. In linguistic studies to date, studies of such discrepancies have been confined mostly to “false friends”, but in terms of near-synonymy across languages, WS shows us that most of that vast ocean of language is composed of hitherto uncharted waters.

⁴ In a recent communication (January 2016) the Sketch Engine team informed me that in WS this initial upper/lower case distinction is to be eliminated.



Works Cited

- Atkins, Sue, and Michael Rundell. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press, 2008.
- Berber-Sardinha, Tony. "Semantic prosodies in English and Portuguese: A contrastive study." *Cuadernos de Filología Inglesa* 9.1 (2000): 93-110.
- Carloni, Giovanna. "Corpora, cognitive styles, English content-specific vocabulary and academic language at university level." *Handbook of Research on Individual Differences in Computer-Assisted Language Learning*. Ed. Mahimi, Mehrak. Hershey PA: IGI Global, 2015. 197-222.
- Dam-Jensen, Helle, and Karen Korning Zethsen. "Pragmatic patterns and the lexical system - a reassessment of evaluation in language." *Journal of Pragmatics* 39.9 (2007): 1608-1623.
- Dilts, Philip, and John Newman. "A note on quantifying "good" and "bad" prosodies." *Corpus Linguistics and Linguistic Theory* 2.2 (2006): 233-242.
- Jakubíček, Miloš, Adam Kilgarriř, Vojtěch Kovář, Pavel Rychlý and Vit Suchomel. "The TenTen Corpus Family." *Proceedings of the 7th International Corpus Linguistics Conference CL 2013*, United Kingdom, 2013. 125-137.
- Kilgarriř Adam and Michael Rundell. "Lexical Profiling Software and its lexicographic applications: a case study." *Proceedings of the 10th EURALEX International Congress 2002*, Copenhagen, 2002. 807-818.
- Kilgarriř, Adam, Pavel Rychlý, Pavel Smrz and David Tugwell. "The Sketch Engine." *Proceedings of the 11th EURALEX International Congress 2004*, France, 2004. 105-116.
- Kilgarriř, Adam, Vojtěch Kovář, Simon Krek, Irena Srdanovic and Carole Tiberius. "A Quantitative Evaluation of Word Sketches." *Proceedings of the 14th EURALEX International Congress 2010*. The Netherlands, 2010. 372-379.
- Kilgarriř, Adam and Iztok Kosem. "Corpus tools for lexicographers." *Electronic Lexicography*. Eds. Sylviane Granger and Magali Paquot. Oxford: Oxford University Press, 2012. 31-55.
- Kilgarriř, Adam, Pavel Rychlý, Vojtěch Kovář and Vit Baisa. "Finding multiwords of more than two words." *Proceedings of the 15th EURALEX International Congress 2012*, Norway, 2012. 693-700.
- Kilgarriř Adam, Vit Baisa, Jan Buřta, Miloš Jakubíček, Vojtěch Kovář, Jan. Michelfeit, Pavel Rychlý, Vit Suchomel. "The Sketch Engine: Ten Years On." *Lexicography ASIALEX* 1 (2014): 7-36.
- Lee, Ching-Ying, and Liu Jyi-Shane. "Effects of Collocation Information on Learning Lexical Semantics for Near Synonym Distinction." *Computational Linguistics and Chinese Language Processing* 14.2 (2009): 205-220.
- McEnery, Anthony, and Richard Xiao. *Corpus-based Contrastive Studies of English and Chinese*. London and New York: Routledge, 2010.
- Munday, Jeremy. "Looming large: a cross-linguistic analysis of semantic prosodies in comparable reference corpora." *Corpus-based Translation Studies: Research and Applications*. Eds. Alet Kruger, Kim Wallmach and Jeremy Munday. London: Continuum, 2011. 169-186.
- Olohan, Maeve. *Scientific and Technical Translation*. London and New York: Routledge, 2016.
- Stewart, Dominic. "Safeguarding the lexicogrammatical environment: translating semantic prosody." *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*. Eds. Allison Beeby, Patricia Rodríguez-Inés and Pilar Sánchez-Gijón. Amsterdam and New York: John Benjamins, 2009. 29-46.
- Stewart, Dominic. *Semantic Prosody: A Critical Evaluation*. London and New York: Routledge, 2010.
- Thomas, James. "Stealing a march on collocation: deriving extended collocations from full-text for student analysis and synthesis." *Multiple Affordances of Language Corpora for Data-driven Learning*. Eds. Agnieszka Leńko-Szymańzka and Alex Boulton. Amsterdam and New York: John Benjamins, 2015. 85-108.
- Whitsitt, Sam. "A critique of the concept of semantic prosody." *International Journal of Corpus Linguistics*, 10.3 (2005): 283-305.
- Xiao, Richard, and Anthony McEnery. "Near synonymy, collocation and semantic prosody: a cross-linguistic perspective." *Applied Linguistics* 27.1 (2006): 103-129.



Dictionaries consulted

Macmillan Dictionary Online: <http://www.macmillandictionary.com/>

Oxford Advanced Learner's Dictionary Online <http://www.oxfordlearnersdictionaries.com/>

Oxford Dictionary Online <http://www.oxforddictionaries.com/?view=uk>

Sabatini Coletti Online http://dizionari.corriere.it/dizionario_italiano/