A Lexicon of Perception for the Identification of Synaesthetic Metaphors in Corpora

Francesca Strik Lievers[1], Chu-Ren Huang[2]
[1]University of Milano-Bicocca, [2]Hong Kong Polytechnic University
francesca.striklievers@gmail.com, churen.huang@polyu.edu.hk

Abstract

Synaesthesia is a type of metaphor associating linguistic expressions that refer to two different sensory modalities. Previous studies, based on the analysis of poetic texts, have shown that synaesthetic transfers tend to go from the lower toward the higher senses (e.g., *sweet music* vs. *musical sweetness*). In non-literary language synaesthesia is rare, and finding a sufficient number of examples manually would be too time-consuming. In order to verify whether the directionality also holds for conventional synaesthesia found in non-literary texts, an automatic procedure for the identification of instances of synaesthesia is therefore highly desirable. In this paper, we first focus on the preliminary step of this procedure, that is, the creation of a controlled lexicon of perception. Next, we present the results of a small pilot study that applies the extraction procedure to English and Italian corpus data.

Keywords: synaesthesia, lexicon, perception

1. Synaesthesia

Linguistic synaesthesia is a metaphorical process of transfer from one sensory modality (source) to a different one (target). A perceptual experience related to one sense is described by lexical means usually associated to a different sense. For example, if we describe a *melody* as *sweet*, we are characterising a hearing-related perceptual experience (*melody*) in terms of taste (*sweet*). *Melody* is the target and *sweet* is the source of the transfer.

Classic studies of synaesthesia are usually based on literary texts. More specifically, many of them focus on Romantic and Symbolist poetry, where synaesthesia finds its most conspicuous and evident expression. Poetic texts include many examples of living synaesthesia, that is, novel associations of sensory expressions (e.g., *sweet light / golden with audible odours exquisite*, Arthur Symons, *The Opium Smoker*, from Ullmann 1957: 275). Interestingly, despite the creativity that characterizes these instances of synaesthesia, several studies (among others, Ullmann, 1957; Shen & Cohen, 1998; Bretones-Callejas, 2001) have shown that synaesthetic transfers tend to be formed following a specific directionality: they have as sources the "lower" senses (touch, smell and taste) and as targets the "higher" senses. For example, *melodious sweetness* is less likely to occur than *sweet melody*. In more recent years, it has been suggested that this tendency may have a cognitive or perceptual motivation (Ramachandran and Hubbard, 2001; Allen-Hermanson and Matey, 2012).

In this study, we focus on non-literary texts. In the majority of cases, synaesthetic phrases that are found in everyday language are instances of conventional synaesthesia, such as *warm voices*, *cold lights*, *sour smells*, etc. (Strik Lievers, 2015b). Data concerning conventional synaesthesia are likely to be more revealing of possible cognitive underpinnings, if compared to living and creative synaesthesia.

A large-scale study based on non-literary texts is still missing. Such a study could either confirm or dismiss the directionality generalisation, and its attached cognitive implications, being a privileged window into the perception / language interface. This paper is a first step in this direction: it presents a method for the identification of instances of synaesthesia in corpora and the results of a small pilot study.

2. How to Find Synaesthesia in Corpora

The first problem that has to be faced in order to conduct a corpus-based study of synaesthesia is its rarity in texts (Marotta, 2012). In order to confirm, dismiss, or make generalizations of any kind, many data are of course needed. Manually searching for such a rare figure would be extremely time consuming, and may not lead to satisfying results. It is mainly in order to cope with the problem of the rarity of synaesthesia that we decided to find a way to retrieve synaesthesia from corpus data in a (semi-)automatic way.

The identification of metaphoric expressions is notoriously not a trivial task. As Stefanowitsch (2006: 1-2) points out: "In the case of metaphor and metonymy, retrieving the relevant data is, at first glance, almost impossible for the simple reason that conceptual mappings are not linked to particular linguistic forms". However, corpus-based studies of metaphor are receiving increasing interest, and many different methods have been proposed for the recognition and resolution of metaphors and more generally of figurative language (among others, Mason, 2004; Deignan, 2005, Stefanowitsch and Gries, 2006; Shutova, 2011; Shutova *et al.*, 2015).

Stefanowitsch (2006: 2-5) recognizes five main approaches to metaphor extraction:

a. Manual searching.
b. Searching for source domain vocabulary.
c. Searching for target domain vocabulary.
d. Searching for sentences containing lexical items from both the source domain and the target domain.
e. Searching for metaphors based on 'markers of metaphor'.

Each approach has pros and cons depending, among

other things, on the type of metaphor to be searched in text. Manual searching (a) is probably the most accurate method, but for obvious reasons it cannot be used to handle a large amount of data. Looking for explicit "markers of metaphor" (e), e.g., *metaphorically*, *literally*, has been shown to be not particularly reliable, and it surely is not in the case of synaesthesia, which is typically not introduced by such markers. Starting from source or target vocabulary (b, c) is the method that has to be followed if, for instance, the aim is to verify how frequent synaesthetic associations are in language.[1] However, if the aim is to understand how synaesthetic associations actually work, no matter whether they are frequent or rare, the best strategy is that of looking for lexical items, from both source and target domain, co-occurring in a single sentence (d). This method is particularly appropriate for synaesthesia since in synaesthesia, unlike in other types of metaphor, the conceptual mappings that are to be found are known in advance: both source and target domain lexical items belong to the domain of perception.

Since conventional instances of synaesthesia, like *sweet voice*, consist in the association of two lexemes referring to two different sensory modalities, we should look for sentences where lexical items referring to two different sensory modalities co-occur. This involves the following three steps: [2]

- Compiling a list of perception-related lexical items, divided by sensory modality.
- Searching for the sentences that include forms of at least two perception-related lexemes from two different sensory modalities.
- Manually checking the extracted sentences, in order to select those that really contain synaesthesia.

In this work, we mostly focus on the first step, on which the success of the automatic extraction procedure crucially depends (Section 3). However, we will also introduce the extraction method and the results reached so far within our pilot study based on English and Italian data (Section 4).

3. Compiling a Vocabulary of Perception-Related Lexical Items

In order to compile a vocabulary of perception-related lexemes, the domain of perception itself has to be circumscribed. How to establish the borders of the lexical field of perception? How many sensory modalities are there? Moreover, is it possible to write an exhaustive vocabulary for this domain?

As is well known, there is no universally agreed view on how many and which sensory modalities there are. Depending on the criteria that are adopted, very small and relatively large systems have been proposed, ranging form three to about thirty senses (Goldstein, 2010). Here we will employ the so-called Aristotelian five-senses system: sight, hearing, touch, smell, and taste. The motivation for this choice is twofold. First, this classification is the most deeply rooted in the cultural background of both English and Italian[3]. Second, our results will be more easily comparable with those described in previous literature on synaesthesia, which is also based on the five-senses classification.

Since we divide the domain of perception into five subdomains, five lists of lexemes have to be compiled. We included verbs, nouns and adjectives, which can all participate in synaesthetic associations, as shown in (1) and (2):

1) *She has a golden*$_{[Adj/Source]}$ *voice*$_{[N/Target]}$
2) *The flowers smell*$_{[V/Target]}$ *sweet*$_{[Adj/Source]}$.

The lists for each sense have been compiled starting from a seed set of basic perception lexemes, obtained partly from introspection, partly from the linguistic literature. This seed set was expanded through available resources. MultiWordNet (Pianta, Bentivogli and Girardi 2001) was used to find troponyms of verbs[4], e.g.:

- *to look* > *to gaze*, *to stare*, *to glance*, etc.

Through the WordSketch function (Rychly, 2008) of SketchEngine (Kilgarriff *et al.*, 2014) we then found the most common direct objects of the verbs. For Italian, this task has been carried out also through Lexit (Lenci, Lapesa and Bonansinga, 2012). Many of these objects are related to the domain of perception as well, e.g.:

- *to smell* > *scent*, *stench*, *perfume*, *odour*, etc.

The resulting list of nouns was merged with the nouns that were already included in the seed set, and further expanded by searching for hyponyms and (near)synonyms in MultiWordNet, e.g.:

- *music* > *tune*, *melody*, *waltz*, etc.

The nouns thus identified were used to find the (sensory)

[1] See Marotta (2012), analysing 16 Italian perception adjectives and the nouns they combine with in corpus data. The study shows that synaesthetic associations are extremely rare in non-literary texts.

[2] This procedure was first presented by Francesca Strik Lievers together with Xu Hongzhi and Xu Ge at the *19th International Congress of Linguists*, University of Geneva.

[3] If our method will be applied to languages spoken by populations that have never been "touched" by Aristotle this choice might need to be reconsidered. To give an example, Ward (2008: 32) reports the following case: "The Cashinahua of Eastern Peru have senses corresponding to skin knowledge, hand knowledge, eye knowledge, ear knowledge, genital knowledge and liver knowledge". Analyzing languages like Cashinahua with "Aristotelian" categories would probably be misleading.

[4] "The troponimy relation between two verbs can be expressed by the formula *To V1 is to V2 in some particular manner*" (Fellbaum, 1998: 79). The notion of troponimy is thus used to classify verbs according to their degree of specificity (as the notion of hyponymy is used for nouns).

adjectives most commonly modifying them, e.g.:

- *taste* > *bitter*, *delicious*, *sour*, *sweet*, etc.

The adjective list was in turn expanded through MultiWordNet.
Some manual post-editing of the first draft of the vocabulary was needed. A few lexemes had to be excluded, due to:

a) Their high polysemy within the domain of perception. The Italian verb *sentire* ('feel'), for example, can refer to many sensory modalities (all except for sight).
b) Their extremely bleached meaning. For example, the English verb *seem* is used as an epistemic verb far more often than as a perception verb.

Only lexical items that clearly have a default meaning linked to exclusively one sensory modality are to be included, otherwise too many false positives would be extracted from the corpus. As a second step, some non-strictly perceptual lexemes had to be added. Lexemes such as *music*, or *colour*, do not describe a perceptual process, but they are closely related to perception through clear conceptual relations, and can therefore participate in synaesthetic associations (e.g., *sweet music*, *soft colours*).
The fact that it is difficult to delimit precisely the domain of perception makes the aim of reaching an exhaustive vocabulary problematic in two ways. First, it is sometimes not fully clear whether a given lexical item should be included or not. For example, does the meaning of the adjective *light* still have some connection with touch synchronically? Second, the list has to be closed at some point. Colours, for instance, may be included; but there are many colours, and getting a complete list could be difficult. In this and similar cases (e.g., musical instruments) we decided to include only the lexemes found through the resources mentioned above. This means that, for example, we have *brown* but not *periwinkle*, because *periwinkle* did not occur in the resources, or it did occur with a very low frequency (it would therefore have few chances to be found in the corpora, and even fewer chances to be found as part of a synaesthetic association).
The result so far is a vocabulary consisting of 434 lexical items for English and 445 for Italian. Both vocabularies are subdivided into five lists, one for each sensory modality:

	Sight	Hearing	Touch	Smell	Taste	tot.
English	120	194	46	28	46	434
Italian	134	149	82	27	53	445

Table 1: Number of lexemes included in our vocabulary of perception

Our vocabularies are relatively small (and - of course - improvable), especially if compared to other resources that have been created recently. To our knowledge, the most comprehensive sensory lexicon is Sensicon (Tekiroğlu *et al.*, 2014). Sensicon includes 22,684 lexemes, together with their degree of association with the five sensory modalities. This is clearly an invaluable resource for many computational applications. However, for the specific purpose of our study, i.e. identifying synaesthetic metaphors in corpus data, this resource might not be suitable. By checking the association scores attributed to items from our lexicon in the Sensicon, we found associations like the following: *melody* displays a higher association with sight than with hearing; *sweet*, *sour*, *acidity*, *taste* display a higher association with smell than with taste; *sticky* shows higher scores for taste, hearing and smell than for touch. These "anomalous" associations would generate errors in the extraction of synaesthesia, thus increasing the amount of manual inspection needed (e.g., *salty taste* would be identified as a synaesthesia, while *bright melody* would not). For our specific task we therefore believe that, at least at this stage, a more controlled lexicon is more suitable. Moreover, Sensicon is only available for English and, to the best of our knowledge, comparable resources for other languages have not been created.

4. (Semi-)automatic Identification of Instances of Synaesthesia

The corpus to be used to search for instances of conventional synaesthesia needs to have at least the following two properties. First, it has to reflect, as much as possible, "everyday language" (e.g., not being a literary corpus or other domain-specific corpus). Second, given the rarity of synaesthesia in text, it has to be large enough to contain a satisfying number of synaesthetic associations. Taking into consideration these two requirements, we chose two web corpora: UkWaC for English, and ItWaC for Italian (Baroni *et al.*, 2009).
For our pilot study, we experimented and compared two different methods for the (semi-)automatic extraction of synaesthesia.

4.1 Method 1

In synaesthesia, a perceptual experience in a given sensory modality is described through lexemes pertaining to another sensory modality. Therefore, our first experiment consists in finding all the sentences that contain at least two lexemes from different lists, i.e., two lexemes pertaining to two different sensory modalities (cf. the "window method" described in Seretan, 2011 for collocation extraction). However, an initial extraction attempt showed that in a sentence window the two perceptual lexemes were seldom synaesthetically connected, as in (3):

3) *Staffed by bright*$_{[SIGHT/Source]}$, *young things who live and breathe music*$_{[HEARING/Target]}$, *they tend to represent clients because they are passionate about their music.*

Although *bright* and *music* refer to two different sensory modalities (sight and hearing), the sentence clearly does

not contain a synaesthesia.
As might be expected, the higher is the distance between the two lexemes, the lower are the chances that they form a synaesthesia. A distance constraint is therefore added: the two perceptual lexemes cannot be separated by more than 1 token. The sentences that are extracted are therefore of the type in (4) and (5):

4) *At last Mortlaok halted and called out in a soft*$_{[TOUCH/Source]}$ *affectionate voice*$_{[HEARING/Target]}$.
5) *It was a disgusting*$_{[TASTE/Source]}$ *sight*$_{[SIGHT/Target]}$*, that bathroom.*

Those in (4) and (5) are good examples of synaesthesia, correctly identified in the corpus. However, a rather heavy manual inspection of the extracted data was needed. The analysis of a sample of 1000 extracted sentences (i.e., sentences potentially containing synaesthesia) for each language shows that those truly containing synaesthesia are 28% in the Italian data and 18% in the English data. In order to improve this result, and thus reduce the load of manual work, an analysis of the errors is needed.

4.1.1 Method 1: Discussion

Some errors have a lexical basis. They are attributable to the polysemy that characterises (a few) words in the lists, which display both perceptual and non-perceptual senses. For example, *ring* is included in the *hearing* list with reference to the sound produced by a bell, but it also appears in the corpus with other meanings, as in (6):

6) *The photographer had retouched the dark*$_{[SIGHT]}$ *rings*$_{[HEARING???]}$ *under his eyes*.

Another example is *reflection*: it can be related to sight (as in *The reflection of light*), but it also has a cognitive meaning, as in (7):

7) *For the North American reader, this portrayal awakens a bitter*$_{[TASTE]}$ *reflection*$_{[SIGHT???]}$

We decided to keep lexemes such as *ring* and *reflection* in the lists because their perceptual meaning is not marginal and, at the same time, their non-perceptual meanings are not very frequent, so that the risk of finding many false positives is not too high.
The main source of errors is however not lexical. In most of the sentences that, after manual inspection, turned out not to contain synaesthesia, the two lexemes from the lists are both used in their perceptual meaning. Despite that, they are not forming a synaesthesia together. A significant number of such errors can be avoided by putting stop words. For instance, if between the two perception lexemes there is a coordinating conjunction, no synaesthesia will be found, as (8) shows:

8) *Enjoy the sights*$_{[SIGHT]}$ and *sounds*$_{[HEARING]}$ *of London*.

However, also adjacent perceptual lexemes can turn out not to be synaesthetically connected, as in (9):

9) *An aromatic*$_{[TASTE]}$ *white*$_{[SIGHT]}$ *flowered herb used as a tonic.*

In this sentence, *aromatic white* is clearly not an instance of synaesthesia: *aromatic* does not modify *white*, it modifies the noun phrase *white flowered herb.*
Is there a way to avoid these "syntactic" errors more efficiently and more effectively? Moreover, is there a way to extract synaesthetic pairs of lexemes separated by more than one token? We will try to answer these questions in Section 4.2.

4.2 Method 2

By analysing the examples of synaesthesia that have correctly been extracted from the corpora, the following part of speech patterns have been detected:

	ItWaC	UkWaC
N Adj / Adj N	97.3%	87%
V Adj / Adj V	2%	11.6%
Adj Adj	0%	0.6%
V N / N V	0.6%	0.6%

Table 2. Part of speech patterns in the instances of synaesthesia extracted from corpora.

In both languages the dominant pattern is "N Adj / Adj N", where the adjective modifies the noun, as in (10):

10) *Must not be missed: a true talent with a golden*$_{[Adj /SIGHT]}$ *voice*$_{[N/HEARING]}$.

In the pattern "V Adj / Adj V" the adjective is connected to the verb as a predicative complement or secondary predicate, as in (11):

11) *Emma looked*$_{[V/SIGHT]}$ *awfully sweet*$_{[Adj/TASTE]}$ *in her bridal things.*

The two remaining patterns are extremely rare. "Adj Adj" is found in our data only in the collocation *red hot*, which is, moreover, a not very convincing example of synaesthesia. As for "V N / N V", the noun is usually the direct object of the verb:

12) *You can savour*$_{[V/TASTE]}$ *the sights*$_{[N/SIGHT]}$.

The data extracted therefore confirm what has often been observed in the literature, namely that in conventional synaesthesia the perceptual lexemes are typically in a dependency relation, which is attributive in most cases (but see also the verb-object relation mentioned above). Therefore, a way to improve the extraction method in

Sentence	Sensory modality		Dep. Type	PoS	Lexeme	
	Source	Target			Source	Target
And the first time they'd kissed he remembered the needy collision of lips and tongues and sticky sweetness	Touch	Taste	amod	N Adj / Adj N	sticky	sweetness
His voice is very bitter but I shake my head and gently touch his hand	Taste	Hearing	nsubj	N Adj / Adj N	bitter	voice
They look sweet!	Taste	Sight	acomp	Adj V / V Adj	sweet	look
Questo albero viene coltivato per i bei fiori dal profumo dolce e per ricavare dalla corteccia una resina dolce commestibile	Taste	Smell	amod	N Adj / Adj N	dolce	profumo
Il Merlot mostra un altro aspetto della sua versatile personalità, con sapori più caldi e speziati.	Touch	Taste	amod	N Adj / Adj N	caldo	sapore
Esistono clarinetti costruiti in metallo e cristallo, poco apprezzati per il loro suono aggressivo e freddo.	Touch	Hearing	amod	N Adj / Adj N	freddo	suono

Table 3. Sample of the database compiled with Method 2

terms of precision is searching exclusively for the syntactic relations that are relevant for synaesthesia. In order to do so, the corpus needs to include also syntactic information. The UkWaC corpus has been tagged with the Stanford Dependency Parser. For Italian, we used the Italian Wikipedia annotated in CoNLL format.[5] Both in the English and in the Italian data we looked for the specific syntactic relations described above.[6] Table 3 shows what the database looks like with English and Italian example sentences.

4.2.1 Method 2: Discussion

Errors due to polysemy are difficult to avoid, as already discussed in Section 4.1.1. They are in most cases due to the inherent polysemy[7] of some lexemes, as in the case of *guitar* (sound / physical object) in (13):

13) *Graeme McDonald, sporting an extremely stylish black*$_{[SIGHT]}$ *and white*$_{[SIGHT]}$ *Danelectro electric guitar*$_{[HEARING???]}$ *was next, and played three songs of pure pop beauty .*

However, the most frequent cause of false positives is attributable to parsing errors. Among them, many are found in sentences where same part of speech lexemes were tagged by the parser as connected by a dependency relation, while being in fact coordinated. The sentence in (14), for instance, has been extracted based on an adjective-noun modification relation between the two perception lexemes *soft* and *yellow*, which are actually coordinated adjectives, both modifying the noun *paste*.

14) *The interior paste of this sort of cheese is soft*$_{[TOUCH]}$ *and straw yellow*$_{[SIGHT]}$ *in colour.*

In future development of this method, it could therefore be helpful to add a part-of-speech constraint. That is, the two perception lexemes must be in a dependency relation *and* they must pertain to different part of speech.

4.3 Discussion and Results

For comparison purposes, the same sample of sentences from UkWaC has been used to extract potential synaesthesiae with the two methods. The results are reported in Table 4.

	Method 1	Method 2
Sentences	19110	19110
Potential synaesthesiae extracted	475	246
"True" synaesthesiae (tokens)	87	139
"True" synaesthesiae (types)	61	101

Table 4. Method 1 and Method 2 compared

Before comparing the results, it is worth noting that sentences containing synaesthesia represent only the 0.5% (Method 1) or 0.7% (Method 2) of sentences in the sample. Although this is just an approximate estimate[8], it

[5] http://wacky.sslmit.unibo.it/doku.php?id=download

[6] The following dependency tags of the Stanford Parser (see de Marneffe and Manning, 2008) are relevant: *amod* (N Adj / Adj N, e.g.: *sweet smell*), *acomp* (V Adj / Adj V, e.g.: *to look sweet*), *dobj* (V N / N V, e.g.: *to savour the sight*), *nsubj* (N Adj / Adj N, e.g.: *his voice is sweet*).

[7] "Inherent polysemy is seen where multiple interpretations of an expression (the nominal head) are available by virtue of the semantics inherent in the expression itself" (Pustejovsky, 2011: 1403). Inherent polysemy contrasts with selectional polysemy: "Selectional polysemy is seen where a novel interpretation of an expression is available due to contextual influences, namely, the type of the selecting expression" (*ibid.*).

[8] It can of course not be guaranteed that *all* existing synaesthesiae have been extracted (this is, at least in part, a

can give an indication of the rarity of synaesthesia in everyday language.

The results of the manual inspection of the potential synaesthesiae extracted shows that Method 2 finds a higher number of "true" synaesthesiae. But what matters more is comparing the ratio between the number of "true" and "potential" synaesthesiae: it is around 18% with Method 1, and around 56% with Method 2. Method 2 is therefore more efficient, significantly reducing the amount of manual revision needed.

Method 2 has a further advantage. While Method 1 can only extract synaesthetic pairs where two perception lexemes are either adjacent or separated by one token, Method 2 does not have a distance constraint. Although in the majority of cases the two lexemes are adjacent or separated by one token, wider distances are also attested, as in the following examples:

15) *The sound*$_{[HEARING/Target]}$ *should be clear*$_{[SIGHT/Source]}$ [Distance: 2]

16) *The songs are wonderful, both awkward and hilarious, and Dan's voice*$_{[HEARING/Target]}$ *is something else, so low and warm*$_{[TOUCH/Source]}$. [Distance: 6]

However, Method 1 is not to be completely discarded, for the simple and "practical" reason that it does not require a dependency-annotated corpus. This can be an advantage especially in view of an extension of the research to a larger number of languages, for some of which there may not be dependency parsers available. If Method 2 is generally to be preferred due to its better performance, Method 1 can therefore be a valid option in some specific cases.

As for the types of sensory associations that have been found, almost every possible combination of senses has been attested. However, in terms of frequency the directionality generalisation is confirmed: most transfers go from the lower to the higher modalities in both English (62%) and Italian (74%) (see Strik Lievers, 2015a for a discussion on directionality).

5. Conclusion

The automatic identification of figurative language is a task that still poses many challenges. However, in the case of synaesthesia, the compilation of a controlled vocabulary of perception-related lexemes has enabled us to obtain encouraging results. A quite heavy component of manual inspection of the extracted data is still needed. However, if we take into account the extreme rarity of synaesthesia, then utility of the described methodology clearly emerges. An (at least partially) automatic procedure to extract synaesthesia from corpora is the only way to obtain enough results for making quantitative considerations, minimizing at the same time manual effort.

consequence of the fact that, as discussed in Section 3, our input lists cannot aim at a 100% coverage of the lexicon of perception).

6. Acknowledgements

We gratefully acknowledge the support of PolyU Project G-YBGM "A Corpus-based Study of Chinese Synaesthesia". We also wish to thank Xu Hongzhi and Xu Ge for their help with the data extraction.

This paper is the outcome of a joint effort. However, for the specific concerns of the Italian Academy, Chu-Ren Huang is responsible for Sections 1 and 5 and Francesca Strik Lievers is responsible for Sections 2, 3, 4.

7. Main References

Allen-Hermanson, S., Matey, J. (2012). Synesthesia. In J. Feiser, B. Dowden (Eds.). *Internet Encyclopedia of Philosophy* (http://www.iep.utm.edu/synesthe/)

Baroni, M. *et al.* (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43(3), pp. 209-226.

Bretones-Callejas, C. (2001). Synaesthetic metaphors in English. *Technical Reports, TR 01-008, International Computer Science Institute*. Berkeley, USA.

De Marneffe, M., Manning, C.D. (2008). *Stanford Dependencies manual*. http://nlp.stanford.edu/software/dependencies_manual.pdf (accessed March 2016).

Deignan, A. (2005). *Metaphor and Corpus Linguistics.* Amsterdam: John Benjamins.

Fellbaum, C. (1998). A Semantic Network of English Verbs. In C. Fellbaum (Ed.). *WordNet. An electronic lexical database*. Cambridge, MA: MIT Press.

Goldstein, E.B. (Ed.) (2010). *Encyclopedia of Perception.* Newbury Park, C. A.: Sage.

Kilgarriff, A. *et al.* (2014). The Sketch Engine: ten years on. *Lexicography* 1(1), pp. 7-36.

Kilgarriff, A., Rychlý, P., Smrz, P. and Tugwell, D. (2004). The Sketch Engine. *Proceedings of EURALEX*, Lorient, France, pp. 105-115.

Lenci, A., Lapesa, G., Bonansinga, G. (2012). LexIt: A Computational Resource on Italian Argument Structure. *Proceedings of LREC 2012*, pp. 3712-3718.

Marotta, G. (2012). Sinestesie tra vista, udito e dintorni. Un'analisi semantica distribuzionale. In M. Catricalà (Ed.). *Sinestesie e monoestesie*. Milano: Franco Angeli, pp. 19-51.

Mason, Z. J. (2004). CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System. *Computational Linguistics* 30(1), pp. 23-44.

Pianta, E., Bentivogli, L., Girardi, C. (2002). MultiWordNet: Developing and Aligned Multilingual Database. In *Proceedings of the First International Conference on Global WordNet*, Mysore, India.

Pustejovsky, J. (2011). Coercion in a general theory of argument selection. *Linguistics* 49(6), pp. 1401–1431.

Ramachandran, V. S., Hubbard, E. M. (2001). Psychophysical investigations into the neural basis of synaesthesia. *Proceedings of the Royal Society of London* B, 268, pp. 979-983.

Rychly, P. (2008). A Lexicographer-Friendly Association Score. In P. Sojka, A. Horák (Eds). *Proceedings of*

RASLAN 2008. Brno: Masaryk University.
Seretan, V. (2011). *Syntax-Based Collocation Extraction*. Dordrecht: Springer.
Shen, Y. and Cohen, M. (1998). How come silence is sweet but sweetness is not silent: a cognitive account of directionality in poetic synaesthesia. *Language and Literature* 7(2), pp. 123-140.
Shutova, E. (2011). *Computational approaches to figurative language*. PhD thesis, Computer Laboratory, University of Cambridge, UK.
Shutova, E., Beigman Klebanov, B., Lichtenstein, P. (Eds.). 2015. *Proceedings of NAACL 2015 Workshop on Metaphor in NLP*. Denver, CO.
Stefanowitsch, A. (2006). Corpus-based approaches to metaphor and metonymy. In A. Stefanowitsch, S. Gries (Eds.). *Corpus-based approaches to metaphor and metonymy*. Amsterdam: John Benjamins, pp. 1-16.
Stefanowitsch, A., Gries, S. (Eds.) (2006). *Corpus-based approaches to metaphor and metonymy.* Amsterdam: John Benjamins.
Strik Lievers, F. (2015a). Synaesthesia: A corpus-based study of cross-modal directionality. *Functions of language* 22(1), pp. 69–94.
Strik Lievers, F. (2015b), Synesthésies: Croisements des sens entre langage et perception. *L'Information grammaticale* 146, pp. 25-31.
Tekiroğlu, S. S, Özbal, G. and Strapparava, C. (2014). Sensicon: An Automatically Constructed Sensorial Lexicon. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1511–1521.
Ullmann, S. (1957). *The Principles of Semantics*. Glasgow: Jackson.
Ward, J. (2008). *The frog who croaked blue. Synaesthesia and the mixing of the senses*. London: Routledge.

Appendix

English lexicon of perception

Sight

(N): *amaranth; beige; black; blue; bronze; brown; carmine; color/colour; coloration; crimson; cyan; eye; gaze; glance; glare; gleam; glimpse; glint; glisten; glitter; glower; gray; green; image; indigo; look; magenta; moonlight; obscurity; orange; photo; picture; pink; purple; red; reflection; scarlet; shade; shadow; shimmer; shine; sight; sunlight; translucency; turquoise; violet; white; whiteness; yellow*

(V): *behold; descry; espy; gape; gawk; gawp; gleam; glint; glisten; glitter; gloat; goggle; leer; ogle; peep; see; shadow; shimmer; shine; stare; watch*

(Adj): *amaranth; amber; beige; black; blackish; blonde; blue; bluish; blurred; bright; brilliant; bronze; bronze; brown; brownish; burgundy; carmine; chromatic; clear; colored; colorless; crimson; cyan; dark; dazzling; dim; dull; emerald; fluorescent; golden; gray; green; greenish; greyish; illuminated; immaculate; indigo; iridescent; magenta; matt; multicolored; nebulous; ocher; orange; pale; pearly; phosphorescent; pink; pinkish; polished; purple; radiant; red; reddish; scarlet; shady; shining; shiny; silvery; solar; suffused; sunny; translucent; transparent; turquoise; vermilion; violaceous; violet; visual; vivid; white; whitish; yellow; yellowish*

Hearing

(N): *accordion; babble; bang; banjo; bass; bell; birr; blare; bleep; boom; brattle; burble; buzz; castanet; cello; chatter; chime; chug ; cithara; clamor; clang; clangor; clank; clap; clash; clink; clop; clump; clunk; concert; contrabass; crack ; crackle; crackling; creak; crunch; deflagration; detonation; ding; dingdong; dong; drone; drum; echo; fiddle; fife; flute; glug; groan; grumble; grunt; guggle; guitar; harmonica; harp; honk ; howl; hum; hustle; hymn; jangle; jazz; jingle; lute; mandolin; melody; meow; moo; murmur; music; noise; noiselessness; orchestra; organ; patter; peal; percussion; piano; pianola; ping; plunk; polka; purl; purr; rattle; ring; ripple; roar; roaring; rumble; rustle; saxophone; screak; scream; screech; shout; silence; sing; sizzle; skriech; snarl; song; sonority; sound; soundlessness; splash; squawk; squeak; squelch; symphony; tambour; thrum; thud; thunder; tick; ticking; ting; tinkle; tootle; trumpet; tune; twang; viola; violin; voice; warble; whir ; whirr; whisper; whistle; whiz ; whizz; xylophone*

(V): *babble; bang; birr; blare; bleep; brattle; burble; buzz; chatter; chime; chug ; clang; clank; clap; clash; clink; clitter; clop; clump; clunk; crack ; crackle; creak; crepitate; crunch; ding; dingdong; dong; drone; eavesdrop; glug; grumble; grunt; guggle; hear; hearken; honk ; howl; hum; jangle; jingle; listen; meow; moo; overhear; patter; peal; ping; plunk; purl; purr; rattle; resonate; resound; ring; ripple; roar; rumble; rustle; screak; scream; screech; sing; skriech; speak; splash; squawk; squeak; squelch; stridulate; talk; thrum; thud; tick; ting; tinkle; tootle; twang; whir ; whirr; whiz ; whizz*

(Adj): *acoustic; audible; bombastic; clarion; deaf; deafening; dumb; guttural; harmonic; harmonious; hoarse; imperceptible; loud; melodic; melodious; noiseless; noisy; querulous; resonant; ringing; silent; sonorous; soundless; squeaky; stifled; strident; tacit; vibrant; vocal; voiceless*

Touch

(N): *cold; coldness; heat; hotness; itch; tickle; touch; warmth*

(V): *fondle; itch; stroke; tickle; touch*

(Adj): *burning; clammy; cold; coriaceous; crisp; dense; dry; fluid; friable; gelatinous; granular; hot; impalpable; incandescent;*

itchy; muddy; oily; piercing; pungent; rigid; rough; sharp; silky; slippery; smooth; soft; spongy; sticky; tactile; tender; unctuous; velvety; viscous; warm; wet; wrinkled

Taste

(N): *acidity; acidulousness; bitterness; flavor/flavour; harshness; salinity; saltiness; sapidity; savor/savour; sourness; sugariness; sweetness; tartness; taste*

(V): *acidulate; savor/savour; sweeten; taste*

(Adj): *acid; acidulous; acrid; appetizing; aromatic; balsamic; bitter; bitterish; bittersweet; delicious; disgusting; fruity; insipid; peppery; refreshing; sapid; smoky; sour; spicy; succulent; sweet; sweetish; syrupy; tart; tasteless; tasty; unsavoury*

Smell:

(N): *fetor; fragrance; malodor/malodour; niff; noseful; odor/odour; perfume; reek; scent; smell; sniff; stench; stink; stinker; whiff*

(V): *perfume; reek; scent; smell; sniff; stink*

(Adj): *fetid; malodorous; odoriferous; olfactory; perfumed; reeking; scented; smelly; stinking*

Italian lexicon of perception

Sight:

(N): *arancione; argento; azzurro; bagliore; beige; bianco; blu; bordeaux; candore; carminio; celeste; color; colorazione; colore; giallo; grigiorosa; indaco; luce; magenta; marrone; nero; occhiata; occhio; ocra; ombra; oro; porpora; riflesso; rosso; scarlatto; sfumatura; sguardo; tenebra; tinta; turchese; verde; viola; violetto; visione; vista*

(V): *apparire; avvistare; brillare; guardare; intravedere; intravvedere; luccicare; mostrare; rilucere; sbirciare; scintillare; scorgere; scrutare; vedere*

(Adj): *abbagliante; ambrato; annebbiato; appannato; arancione; argenteo; argento; azzurrino; azzurro; azzurrognolo; beige; biancastro; bianco; biondo; blu; bluastro; bordeaux; brillante; bronzeo; brunastro; bruno; buio; candido; cangiante; carminio; celeste; chiaro; cieco; colorato; cristallino; cromatico; dorato; fluorescente; fosco; fosforescente; fulgido; giallastro; giallo; giallognolo; grigiastro; grigiorosa; illuminato; incolore; indaco; iridescente; latteo; limpido; livido; lucente; lucido; luminoso; magenta; marrone; multicolore; nerastro; nero; nitido; ocra; offuscato; ombroso; opaco; oro; oscuro; paglierino; pallido; perlaceo; plumbeo; porpora; radioso; riflesso; rosato; roseo; rossastro; rossiccio; rosso; sbiadito; scarlatto; scialbo; scolorito; scuro; sfocato; sfolgorante; sfumato; sfuocato; sgargiante; smagliante; solare; splendido; terso; torbido; trasparente; turchese; velato; verdastro; verde; verdino; verdognolo; vermiglio; viola; violaceo; violetto; visivo; vivido*

Hearing:

(N): *armonica; arpa; baccano; belato; bisbiglio; boato; botto; brusio; cagnara; campanello; canto; canzone; cembalo; cetra; chiasso; chitarra; cicaleccio; cigolio; cinguettio; clangore; concerto; contrabbasso; crepitio; deflagrazione; detonazione; eco; fisarmonica; fischio; flauto; fracasso; fragore; frastuono; fruscio; gemito; gorgheggio; grancassa; grido; grugnito; guaito; jazz; liuto; mandolino; melodia; miagolio; mormorio; muggito; musica; nacchere; nitrito; orchestra; organetto; parola; percussione; pianoforte; pianola; polka; rimbombo; rombo; ronzio; ruggito; rumore; sassofono; scalpiccio; schiamazzo; schiocco; scoppio; scricchiolio; sibilo; silenzio; silenziosità; sinfonia; sonorità; strepitio; strepito; stridio; stridore; suono; sussurro; tamburo; ticchettio; tintinnio; tonalità; tono ; trambusto; tromba; tuono; ululato; urlio; urlo; violino; violoncello; voce; vociare; vocina; xilofono*

(V): *abbaiare; ascoltare; borbogliare; cantare; cigolare; crepitare; fischiare; frusciare; gemere; gorgogliare; gridare; orecchiare; origliare; risuonare; rumoreggiare; scricchiolare; sferragliare; sibilare; squillare; stridere; suonare; tintinnare; trillare; udire*

(Adj): *acustico; armonico; armonioso; assordante; chiassoso; chioccio; fievole; fragoroso; gutturale; melodico; monocorde melodioso; muto; orecchiabile; querulo; roboante; roco; rumoroso; silenzioso; sommesso; sonoro; squillante; stridente; stridulo; udibile; vocale; vocalico*

Touch:

(N): *calore; freschezza; pizzicore; prurito; ruvidezza; tatto; tocco*

(V): *accarezzare; carezzare; graffiare; grattare; lisciare; palpare; sfiorare; strofinare; tastare; toccare*

(Adj): *appiccicaticcio; appiccicoso; ardente; asciutto; bollente; bruciante; caldo; coriaceo; corposo; denso; duro; felpato; floscio; fluido; freddo; fresco; friabile; gelatinoso; gelido; glaciale; graffiante; granuloso; impalpabile; incandescente; infuocato; ispido; lacerante; lieve; liquido; liscio; melmoso; molle; molliccio; morbido; oleoso; ovattato; pastoso; peloso; pruriginoso; pungente; rigido; rovente; rugoso; ruvido; scabro; scivoloso; secco; setoso; soffice; spugnoso; tagliente; tattile; tenero; tenue; tiepido; umido; untuoso; vellutato; vischioso; viscido*

Taste:

(N): *acidità; asprezza; dolcezza; gusto; sapore*

(V): *assaggiare; assaporare; degustare; gustare*

(Adj): *acido; acidulato; acidulo; acre; agro; agrodolce; agrumato; agrumoso; amaro; amaro ; amarognolo; appetitoso; asprigno; aspro; balsamico; disgustoso; dissetante; dolce; dolceamaro; dolciastro; edulcorato; fragrante; ghiotto; gustoso; insipido; mieloso; pepato; piccante; prelibato; salato; sapido; saporito; saporoso; scipito; sciropposo; squisito; stomachevole; succulento; zuccherato*

Smell:

(N): *afrore; aroma ; effluvio; fetore ; fragranza ; lezzo; odore ; olezzo ; profumo ; puzza ; puzzo ; tanfata ; tanfo ; zaffata*

(V): *annusare; fiutare; odorare; olezzare; profumare; puzzare*

(Adj): *aromatico; fetente; fetido; maleodorante; profumato; puzzolente*