Letting in the Light and Working with the Web: A Dynamic Corpus Development Approach to Interpreting Metaphor[1]

Stefano Federici[2] and John Wade[2]

> For, methinks, the Understanding is not much unlike a closet wholly shut from light, with only some little openings left, to let in external visible Resemblences, or Ideas of things without; would the Pictures coming into such a dark Room but stay there, and lie so orderly as to be found upon occasion, it would very much resemble the Understanding of Man.
>
> John Locke (1632–1704)

Introduction

Setting up corpora is a laborious process, requiring time and resources. One problem we may find is that once a corpus has been created, in a very short time its static nature may not reflect the way language is currently used. This raises the question of how to make corpus-building a dynamic process, which may be exploited in a number of ways, including work on dictionaries and grammars, word collocations, sentence structure, semantics and stylistics (Hunston, 2002). Here we focus on a specific aim relevant to our particular area of interest.

This paper sets out to examine the use of a tool designed to facilitate the creation of corpora in small-scale studies of specific language patterns. Such corpora can be used, as in our case in the *Facoltà di Scienze della Formazione* at the University of Cagliari in Italy, to analyse specific aspects of the English language from statistically significant information which can then be used in the production of more authentic and accurate EFL or ESP teaching materials (Hunston, 2002: 99). The object of this experiment is an investigation of how LIGHT is used metaphorically. This choice is based, firstly, on an intuition, i.e. that LIGHT is a very commonly used metaphor underlying the concept of human understanding and perception of the world which surrounds us, as illustrated in the above quotation by John Locke. Secondly, it is a commonly used image in academic educational writing.

Our experiment was carried out using the following procedure. Our starting point is an intuition based on the frequency of the metaphorical use of LIGHT as representing understanding. A series of examples, mostly literary, was first chosen and categorised in order to analyse how LIGHT might be used in metaphor. Secondly, a small sample of LIGHT collocations was collected with a search tool, described in more detail below, designed to "fish for linguistic data" (Sharoff, 2006: 435) on the web. This was followed up by the manual examination and analysis of the collected data. The analysis was then extended through the analogical comparison of the initial manual analysis, allowing the further extraction of a wider sample of data.

In this paper, we will discuss metaphorical uses of LIGHT in a number of different contexts and then illustrate the process of acquiring collocations with LIGHT from the web.

---

[1] Although this paper is the result of a collaborative effort, for the purposes of Italian academic conventions, the following should be noted. John Wade examined metaphorical uses of LIGHT in sections (1) *Light as metaphor*, (4) *Refining the experiment* and (5) *Re-running the experiment*, while Stefano Federici dealt with the design of the tools used in the experiment and the statistical analysis of the data acquired in sections (2) *Tools for textual analysis: the corpus-based approach* and (3) *AI and the corpus-based approach*.

[2] Facoltà di Scienze della Formazione, University of Cagliari, Italy
e-mail: sfederici@unica.it, jwade@unica.it

The crucial step of the acquisition process will then be outlined. This process employs an analogy-based mechanism that extracts examples of figurative usages of LIGHT from the web by discriminating these from literal usages on the basis of analogical similarity to the manual analysis carried out on the initial data collection. Finally our initial intuitive hypotheses are analysed in the light of our findings and the experiment is run again in order to evaluate how a refined hypothesis may be applied to real data.

1. Light as metaphor

John Locke, in a few lines, succeeds in encapsulating what appears to be one of those eternal truths which regard the human capacity for understanding. It is a powerful image, one of those metaphors which seem to underlie our ability to communicate abstract concepts, which in themselves underlie our culture and way of perceiving the world. The play is on light and darkness: light comes into the darkened and disordered room, creating order and rationality in an image of things reflecting the age of the Enlightenment. In this sense the view is, we might say for the moment, a western, European-oriented view of the world.

The origins of this metaphor have deep roots in western culture, and since the focus of this paper is on the English language, our attention is dedicated to English and the concept of light used metaphorically in reference to understanding. This does not, however, exclude further studies into how our approach may be exploited in cross-cultural studies, perhaps in order to understand better those aspects of communication which regard many aspects of human existence.

Our starting point in exploring the theme of the paper is the consideration of two major influences on the English language: William Shakespeare and the 1611 King James Bible (Crystal, 1995: 62-65).

In the first case, Shakespeare's use of language is rich, inventive and makes extensive use of idiom and metaphor which has had lasting repercussions on the use of English in everyday communication today. The following, well-used, example is a case in point:

> But soft! What *light* through yonder window *breaks*?
> Is it the east, and Juliet is *the sun*!
> (William Shakespeare, *Romeo and Juliet*, II/1, 1596)

The beauty of Juliet is compared to the rising sun, illuminating all that surrounds it, a form of vision and, therefore, our first example illustrates how we might take as metaphor (M):

M1 LIGHT IS A VISION

In a second example, we see how LIGHT is exploited metaphorically in two different ways:

> Here *burns* my *candle out*; ay, here it dies,
> Which, whiles it lasted, gave King Henry *light*
> (William Shakespeare, *Henry VI*, Part III, II/6, 1591)

As the mortally wounded Clifford approaches death, his life ebbs inexorably away, represented by the fading light of the candle and, thus, we find the following metaphor:

M2 LIGHT IS LIFE

This is the same light, however, which contributed to making Henry VI great. So LIGHT is both life and greatness. The light of Clifford's life has been given over to the cause of King Henry, contributing to his greatness, which we might interpret as the metaphor:

M3 LIGHT IS HEIGHTENED VISIBILTY

This means that the greatness of Henry is partially due to the 'light' provided by Clifford's burning candle (cf. the adjective 'illustrious', for instance, or the expressions 'He is a shining light in the field of astrophysics' and 'She is a brilliant musician').

If this image is taken further, we realise that, in order to be visible, light is necessary. And here we turn to the second influence mentioned above: the Bible. In the first chapter of *Genesis* in the Old Testament, the creation of the world takes place:

> And God said, Let there be light: and there was light.
> And God saw the light, that it was good: and God divided the light from *darkness*.
> And God called the light *Day*, and the darkness he called *Night*.
> And the evening and the morning were the first day.
> (Genesis 1:3-5, 1611)

It is intriguing to question what significance LIGHT may have in these opening verses of the Old Testament, and how this light is contrasted with darkness in a manner which bears comparison with Locke's conception of how understanding comes about. We might here, for example, interpret light as the divine entity's power, wisdom and knowledge. Indeed, this idea appears to be supported in *Revelation*, where we see once again the divine source of light:

> And there shall be no *night* there; and they need no *candle*, neither light of *the sun*; for the Lord God giveth them *light*; and they shall reign for ever and ever.
> (Revelation 22: 5)

and how this source of light is contrasted, again, with darkness. The metaphor which might be used here is:

M4 LIGHT IS WISDOM

This wisdom represents the divine creative force behind the beginning of things, the means by which order is brought into the world (cf. the expression 'This report shines a light on the iniquities of the judicial system'). By extension, therefore, darkness becomes evil, ignorance and the unknown. Without light we cannot see, and once again light and visibility appear to be inextricably linked together:

> And if the *blind* lead the blind, both shall fall into the ditch.
> (Matthew 15:14)

Thus, a lack of understanding is blindness and ignorance breeds ignorance (cf. the expression 'You've been keeping me in the dark').

Clearly, these are metaphors are highly literary, and such imagery is often associated with more poetic styles of language. J.R.R. Tolkein, for example, adopts an almost biblical style in the example below:

> 'A *darkness* lies behind us', Bëor said; 'and we have turned our backs on it, and we do not desire to return thither even in thought. Westwards our hearts have been turned, and we believe that there we shall find *light*.'
> (J.R.R. Tolkein 1977: 170)

Here, the darkness behind is suffering, violence and conflict, while in the West salvation is to be found. The metaphor might be interpreted as:

M5 LIGHT IS REALISATION

In other words, this is either the process of seeking or reaching a concrete goal or the goal itself.

Remaining in the literary sphere, the constant play on light and dark is also an interesting aspect of Virginia Woolf's works. Her originality lies in examining the feelings, thoughts and memories of the characters of her works at a crucial point in their lives. In *To The Lighthouse*, for example, this regularity or alternation is exemplified by the flashing light of the unreachable Lighthouse. Not only this, but light provides insight into the characters' psyche, a form of self-awareness unexpectedly acquired in a moment of reflection:

> She had been looking at the table cloth, and it had *flashed* upon her that she would move the tree to the middle, and need never marry anybody, and she felt an enormous exultation. (p.191)

During a moment of reflection, Lily receives a 'flash' of inspiration. This is a recurrent theme in the work, bound together by the constant intermittent flashing of the lighthouse, in which the characters seek to make some sense of life in a form of struggle between light and dark. When James, for example, reflects on the past, he attempts to rationalise his feelings:

> What then was this terror, which the past had folded in him, peering into the heart of that forest where *light* and *shade* so chequer each other that all shape is distorted, and one blunders, now with *the sun in one's eyes*, now with a *dark shadow*, he sought an image to cool and detach and round off his feeling in a concrete shape. (p. 200)

Here, both light and shade come into play, as in a painting. Indeed, it is the delicate balance between light and shade in a painting which gives it a sense of depth and perspective, exactly what James is trying to do with his life.

We may draw parallels here with the Joycian concept of 'epiphany' This term often has religious connotations, but if examined from a philological point of view reveals some interesting sources of reflection relevant to this paper. In the Western world, the Greek term 'epiphania' has taken on the meaning of 'the manifestation of Christ to the Gentiles'. This manifestation may be interpreted as a 'coming into the light', i.e. to appear (cf. (i) 'epiphaino' to come into *light*, *shine* forth, *appear* or (ii) 'epiphantos' *visible* or *alive*). Joyce uses the term in the sense of a sudden and, often, unexpected insight is experienced by a character, an almost spiritual, inner understanding of the self. In *A Portrait of the Artist as a Young Man*, for example Stephen Dedalus comes to a sudden realisation:

> Heavenly God! cried Stephen's soul, in an outburst of profane joy.
> He turned away from her suddenly and set off across the strand. His cheeks were *aflame*; his body was *aglow*; his limbs were trembling. On and on he strode, far out over the sands, singing wildly to the sea, crying to greet the advent of the life that had cried to him. (p. 186)

Here the light comes from within, from the soul as expressed in the 'outburst', the burning cheeks, the glowing fire in Stephen's body and a sense of intense ecstasy. Thus, here:

M6 LIGHT IS REVELATION

or a heightened sense of awareness (cf. the expressions 'Her face lit up', 'I've just had a bright idea' or 'I've seen the light').

It is, perhaps, for this reason that such metaphors are often used in educational discourses, where attempts are made to define or describe the process of learning (cf. the expressions 'He's one of the brightest students in the class' and 'It is the teacher's role to

light the way for her students'). Turning from the Western, English speaking world, we find, indeed, that such concepts are not limited solely to the English language. The Russian proverb 'Education is light, lack of it is darkness' (Wade 2006: 108) or, more literally, 'Studying is light, not studying is darkness' (*Wikiquote*, last accessed 25/06/2007) once again employs the contrast between light and dark, knowledge and ignorance. From this viewpoint, the process of education can be considered, if taken from the perspective of Dewey (1938), as the process of guiding the learner towards illumination or enlightenment, an understanding of the world, a process of experience and discovery. Thus, education and learning are viewed as the self-construction of understanding (Wood 1998: 26). In this sense:

M7 LIGHT IS DISCOVERY

By extension, this process can also be seen from a different perspective, that of providing input. The external influence of the teacher, from the point of view of the learner, appears to be demonstrated in a study by Cortazzi and Jin (1999: 168-169), where Chinese students defined good teachers, among other things, as a source of knowledge, using the metaphor 'candle' or 'lamp' and, in a recent informal conversation, a Taiwanese colleague defined the teacher as a 'lighthouse'. Therefore, we also see that:

M8 LIGHT IS A SOURCE OF KNOWLEDGE

Up to now we have seen metaphor used in quite specific contexts. It is, in fact, generally considered that identifying literary or academic metaphor is less problematic than identifying those metaphors we use in everyday language (Wade 2001: 310), since, in particular, the literary metaphor tends to be more marked. Steen (1994: 63), for instance, comments that such metaphors "may jump to the eye" because of their original or poetic nature. Lakoff and Johnson (1980) argue, however, that all language is run through with metaphorical reference, often more subtle and diffuse than we imagine, a kind of semantic framework which makes communication possible:

> We have found [...] that metaphor is pervasive in everyday life, not just in language, but in thought and action. Our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature.
> (Lakoff and Johnson 1980: 3)

In the examples below, we see an illustration not only of how light may be used in terms of some form of 'revelation' in modern, everyday language, but how the light metaphor may also be contrasted with other 'elemental' metaphors such as 'darkness' and 'heat':

> The scientist's job is to *shine light* in the *darkness* [...].
> Ross Anderson, 'We cannot allow the terrorists to terrorise us', *The Guardian*, 20/06/2006

> The debate on special educational needs (SEN) often generates more *heat* than *light*.
> Virginia Bovell, 'Time to spell out the line on special needs', *The Guardian*, 04/07/2006

This is not so far from what we have seen up to this point, either from a syntactic point of view or from the message communicated.

The problem now is to examine our initial hypothesis, based on an intuition supported by examples specifically selected to illustrate the point, by employing a different approach which allows the systematic analysis of a large body of data through the use of specific tools designed for the purpose: acquiring data from the web, categorising and analysing them.

2. Tools for textual analysis: the corpus-based approach

To perform the kind of analysis described above, we developed a tool to acquire word-concordances directly from the web. The tool is a combination of several web/linguistic tools:

- a web spider that acquires a predefined number of web pages
- a tokenizer ("segmenter") that splits acquired web pages
- a rule-based lemmatiser
- a KWIC (*KeyWord In Context*)
- a self-learning analogy-based engine

The web spider (see Figure 1) extracts web pages starting from a given web address. The spider filters out all unneeded web overstructure (HTML tags).

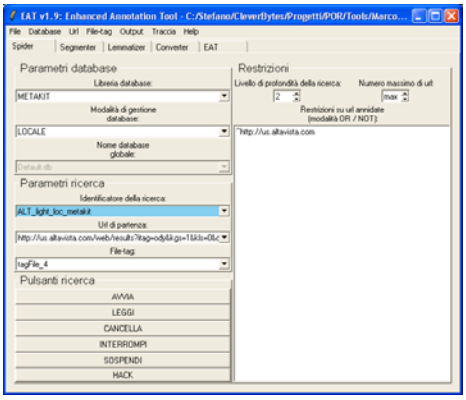

Figure 1: The web spider

Then the lemmatiser associates each word form contained in the extracted web pages to the corresponding lemma. After the corpus has been cleaned and lemmatised, the KWIC will read the corpus by indexing all the lemmas.

The main window of the tool is a KWIC (see Figure 2):

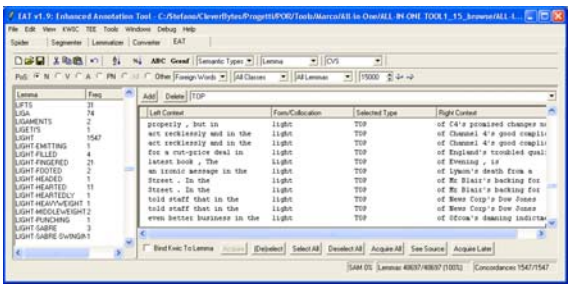

Figure 2: The KWIC tool

The main window of the KWIC tool has two areas (*keyword area*, on the left side; *concordance area*, on the right side) as shown in Figure 3.

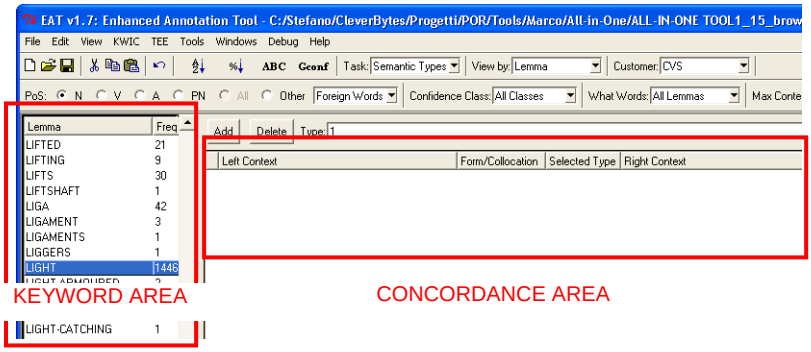

Figure 3: KWIC tool areas

In the keyword area all the words (*lemmata*) contained in the corpus are listed. Then, when a particular lemma is selected in the keyword area, all concordances of that lemma that are contained in the corpus are shown in the concordance area (Figure 4).

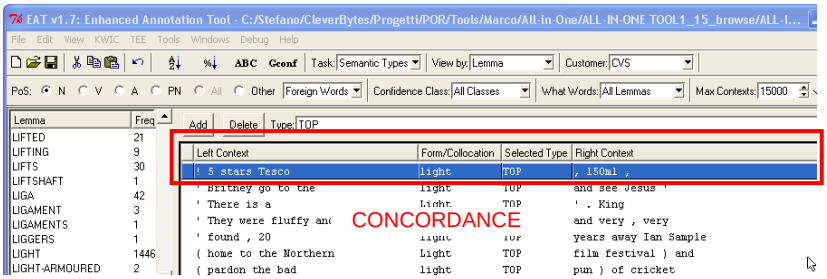

Figure 4: A corpus concordance

When using the tool, the linguist can see in the concordance area all contexts from the corpus collection that contain the word(s) she is interested in and, at the same time, she can tag in a single step all occurrences that have a unique interpretation (*type*).

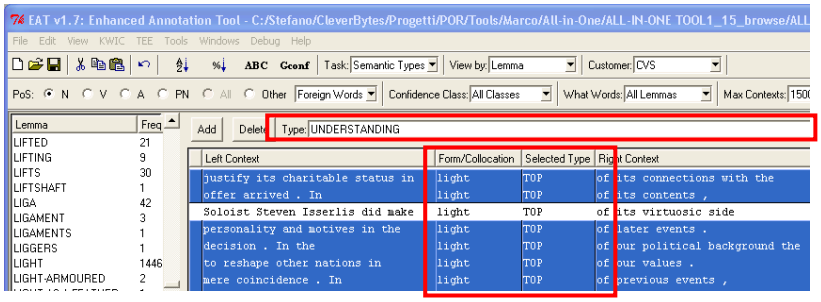

Figure 5: Type assignment in the KWIC tool

This is done by selecting the desired interpretation in the “Type” drop-down menu, or, if not already present in the menu, by typing a new interpretation in the “Type” box (Figure 5. The assigned interpretation will show up in the “Selected Type” column.

Even if the corpus-based approach represents a real improvement with respect to the intuition-based analysis of linguistic data, the above described process still suffers from several limitations:

- *Context search is slow*: browsing all relevant contexts in the corpus (probably thousands of them) is still a very time consuming task, even when it is supported by a KWIC tool.
- *Only a few examples for each interpretation can be listed*: given the relatively slow spidering/browsing/tagging process, linguists can take into consideration only a limited number of contexts from which to derive their analysis. It follows that each analysis of the linguist will be based only on a limited amount of linguistic evidence.
- *Example/analysis coherence*: when the linguist has to browse several thousands of contexts to which to assign the wanted analysis, it is likely that he will assign different analysis to contexts that, even if very similar, are separated in the KWIC window, sometimes, by several hundreds of lines.

3. AI and corpus-based approaches

In this section we are going to show an effective improvement over the corpus-based approach that we illustrated above. The improvement is obtained thanks to a machine learning engine based on machine learning techniques that have been developed for Natural Language Processing and Artificial Intelligence (AI). By using AI techniques we are able to automatise a given task (in this case the assignment of the right interpretation for a given word in a given context) by training the AI engine on a short list of contexts and their manually assigned interpretations[3].

3.1 NLP support

When NLP and AI algorithms are applied to an interpretation task, digital corpora and KWIC are supported by several tools:

- *Tagger*: NLP/AI software to automatically assign the part of speech information to each word in the corpus. By using this tool the KWIC will be able to group together those concordances in which the word LIGHT is a noun.
- *Annotation Tool*: NLP software to manually assign or revise tags that have been automatically assign by the tagger.
- *Machine Learning Engine*: AI software that learns manual annotation performed (by means of the annotation tool) on a small part of the corpus and automatically extends it to the remaining part of the corpus.

[3] For this work we have decided to apply a self-learning engine based on the principles of *paradigmatic analogy* (Federici 1998). This engine has been preferred over other, well-known data-driven techniques developed in IA (such as statistical machine learning, neural nets). However, the choice of such an engine is not a constraint. Whatever technique we are going to use, this will always lead to an effective improvement in the time spent to annotate all extracted contexts with respect to the completely manual assignment of the correct interpretation to each context of the word *light* contained in our corpus. Our choice has fallen on a *paradigmatic analogy* engine as this technique shows good results even when starting from a very small sample of manually annotated contexts (Federici et al. 1996).

By means of these tools a digital corpus (for example a collection of web pages extracted from the Internet) can be prepared and analysed so that its concordances can be shown and processed as shown above.

3.2 The boot-strapping effect

The annotation process described above (manual annotation of a small part of the corpus followed by automatic extension) is a *boot-strapping* process. A boot-strapping process is caracterised by the iterative application of automatic tools starting from a small set of manually analysed data. At each cycle automatically assigned extensions are manually revised (totally or even by sample) so to have a good starting point for the next automatic cycle.

In our experiment 300 contexts out of 900 contexts of the noun *light* from a corpus of 1.5 million words extracted from the "Guardian internet edition" web site were randomly selected. Then we assigned to each of these contexts the correct interpretation of the word *light* among the M1-M8 senses discussed above plus the literal interpretations ILLUMINATION and WEIGHT.

Then the boot-strapping cycle was applied by allowing the AI algorithm to automatically assign one of the possible interpretations to the remaining 900 contexts (*extension*). The automatically assigned interpretations were subsequently manually revised and corrected (if necessary).

3.3 AI algorithms and boot-strapping

To illustrate what the contribution of AI algorithms in a boot-strapping process can be, let us see a real example of how manual annotation performed on several contexts of the word LIGHT randomly extracted from the whole corpus can be fruitfully used to automatically derive the interpretation of LIGHT for the remaining contexts. In Figure 6 four concordances from the initial set of 300 contexts are shown:

```
1.[…] assess his motives in the light/???????????????? of later events […]
2.[…] In light/???????????????? of previous events, however, I'm a tad sceptical […]
3.[…] water flowed through the electric light/????????????? fittings before dripping into
[…]
4.[…] I estimate my home has a hundred electric light/????????????? bulbs, […]
```

Figure 6: Several contexts of 'light' from the acquired corpus

Step 1: manual annotation

In Figure 6 the keyword *light* is shown in bold whereas its interpretations in each context are shown on the right, after the slash sign. As before the first manual annotation step no interpretations have been assigned to the corpus contexts, question marks are shown instead of each interpretation.

In Figure 7 the interpretations that have been manually assigned to the four concordances are shown.

```
1.[…] assess his motives in the light/KNOWLEDGE of later events […]
2.[…] In light/KNOWLEDGE of previous events, however, I'm a tad sceptical […]
3.[…] water flowed through the electric light/ILLUMINATION fittings before dripping into […]
4.[…] I estimate my home has a hundred electric light/ILLUMINATION bulbs, […]
```

Figure 7: Manual annotation step

For the two concordances in position 1 ("[…] assess his motives in the *light* of later events […]") and 2 ("In *light* of previous events, however, I'm a tad sceptical […]") we have manually assigned the interpretation "KNOWLEDGE", whereas to concordances 3 ("[…] water flowed through the electric *light* fittings before dripping into […]") and 4 ("[…] I estimate my home has a hundred electric *light* bulbs, […]") we have assigned the interpretation "ILLUMINATION".

Step 2: automatic extension

In **Error! Reference source not found.** two concordances drawn from the set of the ones that have not been manually annotated are shown.

Manually annotated concordances

```
1.[…] assess his motives in the light/UNDERSTADING of later events […]
2.[…] In light/UNDERSTADING of previous events, however, I'm a tad sceptical […]
3.[…] water flowed through the electric light/ILLUMINATION fittings before dripping into […]
4.[…] I estimate my home has a hundred electric light/ILLUMINATION bulbs, […]
```

New concordances

```
1.[…] single older people who fear for their future alone in the light/?????????????? of
these events […]
2.[…] I think it's just possibly a few spots of very light/???????????? rain  […]
```

Figure 8: New concordances

On the basis of what has been learned from the task of assigning the correct interpretation to the set of initial concordances, i.e. by *extending* the manually assigned interpretation, all the remaining concordances in the corpus are automatically annotated.

Manually annotated concordances

1.[…] assess his motives in the **light/KNOWLEDGE** of later events […]
2.[…] In **light/KNOWLEDGE** of previous events, however, I'm a tad sceptical […]
3.[…] water flowed through the electric **light/ILLUMINATION** fittings before dripping into […]
4.[…] I estimate my home has a hundred electric **light/ILLUMINATION** bulbs, […]

New concordances (*Automatic extension*)

1.[…] single older people who fear for their future alone in the **light/KNOWLEDGE** of these events […] **(OK)**
2.[…] I think it's just possibly a few spots of very **light/ILLUMINATION** rain […] **(NO)**

Figure 8: Automatic extension

The automatic extension step, based on the AI algorithm, assigned two different interpretations to the new concordances. The chosen interpretations ("KNOWLEDGE" and "ILLUMINATION") are shown in Figure 9.

The interpretation that has been automatically assigned to the first new concordance ("KNOWLEDGE") is clearly correct. As to the second new concordance, the interpretation "ILLUMINATION" is instead wrong. The correct interpretation should be "WEIGHT".

How does automatic extension work?

In Figure 9 we can see the most important elements that contributed to assigning the "KNOWLEDGE" and "WEIGHT" interpretations to the new concordances.

Manually annotated concordances

1.[…] assess his motives in the **light/KNOWLEDGE** of later events […]
2.[…] In **light/KNOWLEDGE** of previous events, however, I'm a tad sceptical […]
3.[…] water flowed through the electric **light/ILLUMINATION** fittings before dripping into […]
4.[…] I estimate my home has a hundred electric **light/ILLUMINATION** bulbs, […]

New concordances (*Automatic extension*)

1.[…] single older people who fear for their future alone in the **light/KNOWLEDGE** of these events […] **(OK)**
2.[…] I think it's just possibly a few spots of very **light/ILLUMINATION** rain […] **(NO)**

Figure 9: Hinges of automatic extension

The selection of "KNOWLEDGE" as the correct interpretation of the first new concordance has been driven by the words "in… of… events" that are found both in the manually annotated contexts and in the new one. Instead, the second new context for *light* did not contain the word "electric", which was contained in manually annotated concordances 3 and 4. Thus, the occurrence of word *light* in the second concordance has been interpreted as "ILLUMINATION" on the basis of elements that are not visible in the limited contexts shown in the concordance, possibly even for a complete lack of contextualising elements. This case is not rare when we start with such a small number of manually annotated examples, as is the case in our experiment.

Step 3: manual revision

The last stage of our boot-strapping cycle is the manual revision step. During the manual revision step all (or part of) automatically assigned interpretations are manually revised (Figure 12). The first new concordance will be then confirmed without the need for any action by the linguist. The second new concordance will have instead to be corrected by selecting the correct interpretation. This task is really fast when performed by means of the KWIC Annotation Tool: the correct interpretation can be indeed selected from a drop-down menu (Figure 5 and Figure 10). Keyboard shortcuts are also provided.

New concordances (*Manual revision*)

1.[…] single older people who fear for their future alone in the **light**/KNOWLEDGE of these events […] (OK)
2.[…] I think it's just possibly a few spots of very **light**/***WEIGHT*** rain […] **(OK)**

Figure 10: Manual revision step

3.4 Results

Does reviewing a (partially) automatically annotated list of contexts improve on manually selecting all relevant examples for all given interpretations from a corpus? Our experiment answered a clear "yes" to this question. Indeed, the total amount of time needed to perform the 3 steps of the boot-strapping process was about 5 hours. On the contrary, the full manual annotation of 900 contexts took about 7 hours. By comparing these figures we can see that we saved (7 – 5) / 7 = 29% of time. In this particular case, about 2 hours of manual work.

Result evaluation

We think that results from this experiment are promising and they are worth a deeper analysis. Some limitations that we forcefully imposed on this experiment can indeed be easily removed when this process is to be carried out by a professional linguist.

First, taking into account only 300 contexts in a machine learning task for linguistic analysis is not the best strategy. To reliably identify analogies among all the contexts bearing the same interpretation a greater number of correct contexts (i.e. manually annotated or manually revised) is needed. We must remember indeed that some of the 10 possible interpretations are present in the corpus with a low frequency. For example, some interpretations were assigned in the manual annotation step to less than 2% of the 300 selected contexts. Then, for some interpretations, less than 6 contexts were available to the machine learning algorithm, a very low amount even for a *paradigmatic analogy* algorithm.

Second, applying a boot-strapping algorithm allows us to make use of a greater number of cycles (instead of the single cycle we used in our experiment). This strategy allows us to move a significant number of contexts from the manual (heavier, as there are no suggestions) annotation step to the manual (lighter, as based on automatically assigned suggestions) revision step. It is reasonable to think that a boot-strapping process based on a greater number of automatic extension/revision steps, for example with eight boot-strapping cycles instead of just one, could give better results in terms of time-saving.

A few simple calculations give us some predictions about the advantages that this more granular boot-strapping process can achieve. As manual annotation of 900 contexts

took 7 hours, it follows that, on the average, 0.75 hours are required for every 100 contexts[4]. Our boot-strapping experiment took 5 hours, several of which were spent to manually annotate the first sample of 300 contexts. Then the manual annotation step required about 2.25 hours (0.75 hours for each one of the three samples of 100 contexts). It follows that manual revision took about 2.75 hours (5 hours minus 2.25 hours). Then, on the average, 0.45 hours for every 100 contexts. To sum up, the time required for the 8 boot-strapping cycle experiment can be estimated as about 4.35 hours (0.75 hours for the manual annotation step + 0.45 hours times 8 boot-strapping cycles). The estimated gain is now (7 – 4.35) / 7 = 38% of time.

A further consideration says that when the boot-strapping cycle is repeated more than once, at each new cycle the percentage of correctly extended interpretations increases. Then, the time needed to perform the manual revision step will go down accordingly. Indeed, in our experiment the percentage of correctly assigned automatic extensions, after the first cycle, was 49%, thus requiring a manual revision of the interpretation automatically assigned for more than 50% of the remaining 600 contexts.

This accuracy rate is not very high when considered in isolation. However, to understand the real value of this accuracy rate, we can compare it with several baselines given by simpler strategies, such as i) random selection or ii) selection of the most frequent interpretation (M8, "LIGHT IS A SOURCE OF KNOWLEDGE"). Comparison of the 3 strategies is shown in **Error! Reference source not found.**.

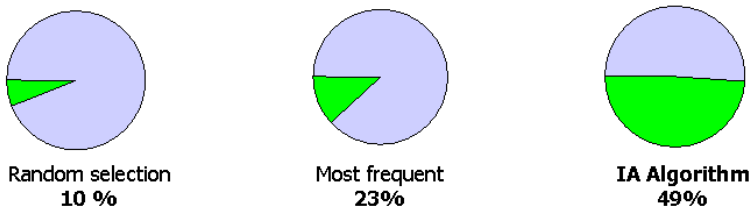

Figure 12: comparison of automatic extension strategies

By looking at the comparison chart, we see that neither the random selection strategy nor the selection of the most frequent interpretation will get results comparable with the application of a simple IA strategy. Indeed, random selection of one of the 10 possible interpretations cannot get further than 100 / 10 = 10% correct extensions. Alternatively, selection of the most frequent interpretation cannot get further than 23% of correctness. In this respect, a correctness rate of 49% achieved without tailoring the machine learning strategies to the specific task at stake can be considered a significant result.

4. Refining the classification

From our initial analysis it appears clearly that metaphorical uses of LIGHT are quite common. However, a further classification would also seem to be necessary to provide more statistical data on such uses.

By analysing the examples automatically extracted by the AI algorithms, they seem to point towards an underlying metaphor, that of visibility, which gives rise to interpretations at other levels of 'abstractness'. Therefore we may start from a first hypothesis (HM):

[4] This estimate is a lower bound. Indeed we can suppose that manual annotation will speed up as annotation goes on. This thanks to the experience acquired by the linguist during the manual annotation process.

HM1 LIGHT IS SEEING

(cf. the expressions 'I don't see what you mean' or 'Do you see my point?') and extend this concept towards distinct perspectives which we may define as:

HM2 LIGHT IS AN INTERNAL OR SPIRITUAL REVELATION
HM3 LIGHT IS AN EXTERNAL, INSPIRATIONAL INFLUENCE

It is clear that the examples given above are based on a further intuition, which brings to mind Samuel Johnson's monumental, if at times quirky, *Dictionary of the English Language* (1755). In an anecdotal quote, he provides an illuminating insight into the problem here discussed:

> We all *know* what light is; but it is not easy to *tell* what it is.

Nevertheless, it is argued here that with sufficient statistical data we may, indeed, begin to 'tell what light is' and it was decided to re-run the experiment with our new hypotheses.

5. Re-running the experiment

To test the new hypothesis of classification we re-ran the experiment with exactly the same configuration but we used only the 5 possible interpretations described above (HM1, HM2, HM3, ILLUMINATION, WEIGHT). This time we had an increase in the correctness rate that reached 57%, instead of 49% reached in the previous experiment. The fact of just having fewer interpretations fails to completely justify this improved performance. Indeed, we can notice that among the automatically assigned interpretations there is a superior consistency.

Therefore, to conclude, we examine now a selection of examples that have been automatically assigned by the AI engine, which serve to illustrate some metaphorical uses of LIGHT in specific contexts. These choices are <u>not based on frequency</u>, but on the particular collocations of LIGHT according to meaning.

Structures with LIGHT as a source of 'internal revelation' (cf. HM2 LIGHT IS AN INTERNAL OR SPIRITUAL REVELATION) is relatively rare in the corpus (Table 1):

[…]has now seen the **light** about the need for[…] […]time to let in the **light**[…]
Table 1: HM2 REVELATION/INTERNAL

Instead, LIGHT as an external influence (cf. HM3 LIGHT IS AN EXTERNAL, INSPIRATIONAL INFLUENCE), focusing attention on something, is seen far more frequently with variations on 'shed light on', 'in the light of' and 'come to light' (Table 2):

[…]shine the **light** of international scrutiny on human rights[…] […]of modern Britain shone the **light** of understanding into many[…] […]casts **light** on dark childhoods[…] […]living fades in the harsh **light** of reality[…] […]in the **light** of the economic data[…] […]considers him in a new **light**[…] […]viewed in a whole different **light**[…] […]seen in this **light**[…] […]in a less than flattering **light**[…] […]The incident only came to **light** relatively recently[…] […]have not been brought into the **light**[…] […]pulls family skeletons into the **light**[…]
Table 2: HM3 INSPIRATION/EXTERNAL

The light can also come from a source, a form of 'heightened visibility' (cf. M3) which through its strength has an influence on those around (Table 3):

[…]a leading **light** of British pop[…] […]can be our guiding **light**[…]
Table 3: HM3 INSPIRATION/EXTERNAL

or may serve as a fount of wisdom and knowledge (cf. M4 and M8) (Table 4):

[…]good teachers **light** candles in dark places[…]
Table 4: HM3 INSPIRATION/EXTERNAL

Finally, a close examination of the data collected reveals that there are cases where LIGHT draws people towards it as it were a goal or a hope, as in M5 LIGHT IS REALISATION, which does not fall easily into the categories HM2 and HM3 and should therefore be classified in a category of its own. This category is not so clearly classifiable as UNDERSTANDING, since enlightenment comes about once the light has been reached (Table 5):

There is some **light** on the horizon[…] […]provide **light** at the end of a very dark tunnel[…] […]have seen a flicker of **light** in the gloom[…] […]opens up a chink of **light** for Henman[…] […]seekers after a new **light**[…]
Table 5: M5 REALISATION/EXTERNAL

Considering all these examples together, we can see that the question is rather more complex than first appeared and that rather than UNDERSTANDING itself, LIGHT might be taken as 'the means by which understanding is brought about'. Thus, further work will need to be carried out in order to refine our findings.

6. Conclusions

At this point we can begin to examine the advantages and disadvantages of creating linguistic resources using the strategies illustrated in this paper. An apparent disadvantage in

applying this strategy is the time necessary to study and implement an AI algorithm which is able to deal with the information contained in an annotated corpus and the integration of this engine with a linguistic tool which allows the visualisation and annotation of the corpus contexts (KWIC tool) and then automatically annotate the rest of the corpus. This disadvantage is quite obvious, since softwares for the management and automatic analysis of linguistic data are widely available (Baroni, 2006), even with open-source licences, which allow a satisfactory personalisation and easy integration with specialised algorithms, such as those used in this experiment.

The advantages of this approach can be found in the use of a strategy which allows the manual annotation of a part of the relevant contexts in the corpus, giving the linguist greater speed in the selection of examples of interest in the study being carried out. As a consequence, a larger collection of examples derived from authentic language use can be used to create a linguistic resource which might be made available in electronic format. Finally, the results obtained from the meanings automatically selected by the AI engine allow the linguist to avoid problems of interpretation in cases where meanings or usage might be ambiguous or unclear, as in the distinction between figurative and literal language.

The use of automatic or semiautomatic tools to access corpora of authentic texts in a digital format supported by AI strategies allows the construction of linguistic resources of higher quality and in significantly shorter time than introspective strategies, the direct consultation of printed texts or the use of software tools for access to resources in digital format. The greater speed and quantity of authentic linguistic data available on the web provides a source of useful data in a reasonably limited amount of time. The procedure outlined in this work can certainly be refined by extending the corpus used and, taking the object of this study, examining how variations on LIGHT, for example, 'luminosity', 'illumination', 'brightness', 'vividness', 'brilliance', might be analysed from a metaphorical point of view.

The promising results of this experiment open the way to the further development of the tools described here in order to facilitate small-scale, ad hoc studies which might have subsequent practical applications, which in our case includes the teaching of English in specific contexts.

References

Baroni, M. (2006), *Tools and Resources*, available from http://sslmit.unibo.it/~baroni/tools_and_resources.html

Cortazzi, M. and L. Jin (1999) 'Bridges to learning: metaphors of teaching, learning and language', in L. Cameron and G. Low (eds) *Researching and Applying Metaphor*, Cambridge: Cambridge University Press.

Crystal, D. (1995) *The Cambridge Encyclopedia of the English Language*, Cambridge: Cambridge University Press.

Dewey, J. (1938) *Experience and Education*; Kappa Delta Pi.

Federici, S. (1998) 'An efficient algorithm for the automatic building of a lexicon from corpora', in *Proceedings of EURALEX 98*, Liège, Belgium.

Federici, S., V. Pirrelli e F.Yvon (1996), 'A dynamic Approach to Paradigm-driven Analogy', in S. Wermter, E. Riloff e G. Scheler (eds) *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, Springer.

Gola E. (2005) *Metafora e Mente Meccanica*, Cagliari: CUEC.

Hunston, S. (2002) *Corpora in Applied Linguistics*, Cambridge: Cambridge University Press.

Jin, L. and M. Cortazzi (1998) ‘The culture the learner brings: a bridge or barrier?’, in M. Byram and M. Fleming (eds), *Language Learning in Intercultural Perspective*, Cambridge: Cambridge University Press.

Joyce, J. (1916/1992) *A Portrait of the Artist as a Young Man*, Penguin: London.

Lakoff, G. and M. Johnson (1980) *Metaphors We Live By*, Chicago: University of Chicago Press.

Locke, J. (1690/1975) *An Essay Concerning Human Understanding*, Oxford: Oxford University Press.

Shakespeare, W. (1996) *The Complete Works of William Shakespeare*, Wordsworth: Ware.

Sharoff, S. (2006) ‘Open-source corpora: using the net to fish for linguistic data’, in *The International Journal of Corpus Linguistics* 11/4, pp. 435–62.

Steen, G. (1994) *Understanding Metaphor in Literature*, Harlow: Longman.

Tolkein, J.R.R. (1977) *The Silmarillion*, London: GraftonBooks.

Wade, J.C. (2001) ‘Exploring language through literature’, in *Annali della Facoltà di Scienze della Formazione dell’Università di Cagliari* (Nuova Serie) XXIV, pp. 297–321.

Wade, J.C. (2006) *English for Education*, Venezia: Cafoscarina.

Wikiquote available from http://en.wikiquote.org/wiki/Russian_proverbs (last accessed 25 June 2007).

Wood, D. (1998) *How Children Think and Learn* (Second Edition), Blackwell: Oxford.

Woolf, V. (1927/1992) *To The Lighthouse*, Penguin: London.