University Centre for Computer Corpus Research on Language Technical Papers

Volume 18 Special issue. Proceedings of the Interdisciplinary Workshop on Corpus-Based Approaches to Figurative Language 27 March 2003

held in conjunction with the Corpus Linguistics 2003 conference

Editors: John Barnden, Sheila Glasbey, Mark Lee, Katja Markert and Alan Wallington

LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER
LANCASTER

UCREL

UCREL
Computing Department
Lancaster University
Lancaster
LA1 4YR
United Kingdom
Phone: (+44) 1524 593802
Fax: (+44) 1524 593608
Email: ucrel@lancaster.ac.uk

ISBN 1-86220-147-1

Further copies may be obtained from
http://www.comp.lancs.ac.uk/ucrel/tech_papers.html

Proceedings of the Interdisciplinary Workshop on Corpus-Based Approaches to Figurative Language
Thursday 27th March 2003 as part of
CORPUS LINGUISTICS 2003
Table of Contents

Forward

Figurative language is pervasive in all kinds of discourse and as a phenomenon has attracted considerable interest from a wide variety of fields including linguistics, psychology, artificial intelligence and philosophy. However, the majority of work has been guided by linguistic intuition and not analysis of real usage. We intended this workshop to address this by focussing on the use of corpora to investigate figurative language. The following areas were of particular interest:

- corpus-based studies of figurative aspects of any language
- corpus-based studies of polysemy and context-sensitive meaning, in their relation to figurative language
- multilingual or cross-lingual studies of figurative language
- computational models of figurative language interpretation or generation, using results from corpora for guidance or being substantially evaluated on corpora
- psychological models of figurative language processing, using results from corpora as a significant contribution
- relationships between processing models and corpus studies

A second intention of the workshop was to explore the methodological issues of using corpora to study figurative language, such as:

- illumination of the concepts of literalness, metaphor, metonymy etc. through corpus studies
- interannotator agreement on what constitutes figurative language, metaphor, metonymy etc.
- specific linguistic cues for figurative language, including studies of their frequencies and reliability and evaluation of their amenability to automated detection
- corpus design and corpus analysis tools for figurative language studies
- effects of domain, genre or corpus type on studies of figurative language (including cross-corpus studies)

The workshop was a follow-up to our previous workshop at Corpus Linguistics 2001. A second workshop was felt appropriate since in the last two years there has been a great deal of interest and, as the papers in this proceedings show, new and exciting research in this area.

John Barnden, University of Birmingham
Sheila Glasbey, University of Birmingham
Mark Lee, University of Birmingham
Katja Markert, University of Edinburgh
Alan Wallington, University of Birmingham

A diachronic approach to figurative language

Kathryn Allan
Department of English Language
University of Glasgow
K.Allan@englang.arts.gla.ac.uk

Introduction

Metaphor studies have long been challenged by questions about the nature of metaphor, and even after many centuries of study there is surprisingly little consensus about what actually constitutes metaphor. Dictionary definitions of the term vary, and many would be disputed by cognitive linguists. The *Oxford English Dictionary* (in a revised 3rd edition entry) defines metaphor as "A figure of speech in which a name or descriptive word or phrase is transferred to an object or action different from, but analogous to, that to which it is literally applicable"; the *American Heritage Dictionary* offers "A figure of speech in which a word or phrase that ordinarily designates one thing is used to designate another, thus making an implicit comparison". Definitions like these rightly reflect widely-held popular beliefs about metaphor, but within cognitive linguistics these beliefs have been disputed and discredited in recent research – metaphor is no longer regarded as a figure of speech only and has been shown to be common and pervasive, and theories that metaphorical mappings are based on similarity or comparison have been rejected as inadequate or simply mistaken. However, it seems to me that there is not yet any widely agreed alternative to this kind of definition within linguistics itself. Recently, with the increasing interest in electronic corpora and artificial intelligence, there have been renewed efforts to find some reliable procedure for identifying metaphor, and at the heart of this must be a generally acceptable definition of metaphor.

A further complication in the debate is the existence of metaphors regarded by many as conventionalized to the extent that they 'die' or cease to be metaphorical. Work in cognitive linguistics, concentrated on system-wide "metaphors we live by" (Lakoff & Johnson 1980), has diverted much attention away from this issue by shifting focus to the cognitive mechanisms that underlie metaphorical mappings, but there is still some uneasiness about the difference between more and less 'active' metaphors.

By taking a diachronic approach to metaphor, I would contend that it is possible to side-step these issues and adopt a pragmatic, data centred stance. My analysis of the target concept INTELLIGENCE starts from an examination of the etymological development of a group of lexical items, to identify earliest meanings and stages in semantic change, and this approach renders it unnecessary to draw up any strict guidelines for metaphor until these can be based on evidence. I hope that this will preclude a situation in which one begins with a rule that proves to be prohibitively theoretical and narrow, and which has to be supplemented to deal with 'anomalous' real examples. Issues of metaphoricity and conventionality in particular mappings also become largely irrelevant, since the important point for my study is the metaphorical basis of meaning change and the processes on which this change depends.

The data I have used is from the *Historical Thesaurus of English*, an ongoing project at the University of Glasgow, which presents lexical items from Old English to Present Day English (taken from the *Oxford English Dictionary* and the *Thesaurus of Old English*) chronologically and by semantic field. This resource offers an unusual opportunity for historical study, and in particular for metaphor study, since it presents areas of vocabulary grouped by concept far more comprehensively than any previous publication. My observations are based on a corpus of 1076 HTE entries, made up of 465 nouns and 611 adjectives. These entries are dated from OE to PDE; 119 entries, just over 11% of the total data, date as far back as OE, and around 40% are words that are considered current (although a number of these words are archaic, rare or in specialised usage). This corpus is stored in a very simple Access database, which allows searches by various criteria including part of speech, date, concept etc.

Although my study is not intended to be quantitative, I have used quantity as a basic indication of the source fields that are particularly productive and therefore characterise our conceptualisation of intelligence. In this paper I will present the three most quantitatively important source concepts I have identified within the INTELLIGENCE data, the SENSES, ANIMALS and DENSITY, and show that each group raises particular issues about the nature of metaphor. I have presented a sample of 20 simplified entries from each group to give an indication of the nature of the data.

SENSES

meaning	core concept	entry	part of speech	date
clever	VISION	**gleaw**	aj	OE
stupid	VISION (LIGHT)	**dwæs**	n	OE
stupid	VISION	**unwitty<unwittig**	aj	OE-1670+1859
clever	VISION	**wise<wis**	aj	OE>
clever	TASTE	**sage**	aj	1297-(a1872)
clever	VISION	**goky**	n	1377
clever	TOUCH	**perceiving**	aj	c1410-1645
stupid	HEARING	**deaf**	aj	c1440-1482
clever	VISION	**clear-eyed**	aj	1530>
clever	TOUCH	**of a far fetch**	aj	1574
stupid	VISION (LIGHT)	**unilluminated**	aj	1579>
clever	TOUCH	**conceitful**	aj	1594-1607
stupid	VISION	**woollen-witted**	aj	c1600-1635
clever	VISION	**wiseling**	n	1633-1765+1914>
stupid	TASTE	**insipid**	n	a1700-a1834
stupid	HEARING	**dunny**	n	1709
clever	TOUCH	**clever**	aj	1716>
clever	VISION (LIGHT)	**bright**	aj	1824+1885>
clever	TOUCH	**tactful**	aj	1864>
clever	TASTE	**savey/savvy**	aj	1905>

The senses have long been recognised as integral to our conceptualisation of mental perception itself – vision and touch, particularly, are pervasive in our vocabulary about knowing and understanding, and this in part accounts for the amount of research into the connection between them that has been undertaken within a variety of disciplines. Nearly a fifth of the words included in my database have connections with the senses[1], and of these, around 70% are used to signify cleverness. This is markedly different from the balance in the data as a whole, in which far more entries signify stupidity than cleverness, but presumably it can be explained by a focus on the senses as conduits of knowledge rather than on a lack of the senses as an impediment to cognition.

All of the senses except SMELL are represented in the data, with a huge bias towards VISION, as shown below. I have included TOUCH in the SENSES group, although it should be pointed out that almost all of the items relating to this concept are more specifically connected with grasping. I would contend that this is a special case of touching which incorporates the concept of possession or enclosure.

concept	**entries**	**% of SENSES data**	**% of total data**
SENSES	204	100	18.96
VISION	158	77.45	14.68
TOUCH (GRASP)	32	15.69	2.97
TASTE	12	5.88	1.12
HEARING	6	2.94	0.56
SMELL	0	0	0

1. I have classified the data by 'core concept': this is a purposely general term, since it includes metaphorical sources but also concepts such as those represented by one element in a compound word that might be regarded as more 'literal' in motivation, for example 'brain'. Included in the core concept 'senses' is LIGHT, which I regard as a special extension of VISION (this is discussed below). I have also identified a further group of data, the core concept of which I have termed SENSE/FEELING, but this has not been included here. Since words within this particular group are not related to particular physical senses (eg VISION or TOUCH), it is difficult to determine whether they can correctly be associated with the physical senses or are more sensibly identified with some kind of abstract 'mental' sense (or, as seems most likely, whether they carry a generalised meaning with elements of both).

It should also be pointed out that words can be classified with more than one core concept if they have undergone significant meaning shifts. For example, words derived from the Latin root *capere*, such as perceived aj c1400, appear both in SENSE-TOUCH, following the meaning of this root, and in SENSE-VISION, reflecting a semantic shift.

As Sweetser points out (1990:39), the importance of the senses and especially vision as source concepts for intelligence is easily understood, since our knowledge about the world is based on information gained through the senses. In this respect, this is a textbook case for Conceptual Metaphor Theory, since it demonstrates the way in which our physical being cannot be separated from the way we conceptualise, and consequently affects language. From very early experience, humans have access to knowledge and understanding through the physical senses, and as a result the process (gaining knowledge/understanding) and the end result (being knowledgeable/having understanding) are inextricably linked, to the extent that one affects the way the other is perceived. In other words, the link is made involuntarily, and the resulting connection is classified as a metaphor. This is consistent with two complementary theories that have been proposed recently, which form part of the Integrated Theory of Primary Metaphor proposed by Lakoff & Johnson (1999:46ff). The first of these is Grady's theory of primary metaphor, and in his list of proposed primary metaphors he identifies both vision and grasping as sources (Grady 1997:296-7). The second, Johnson's theory of conflation, concerns the way in which concepts are initially acquired (Johnson 1999). Johnson looked at the way vision vocabulary is learned and used, and concluded that for the young child *look* and *see* have conflated senses that incorporate both physical vision and mental perception. This is a significantly different claim from Metaphorical Acquisition Hypothesis, which suggests that metaphorical links are made when one (probably concrete) meaning is learnt and then transferred to another (probably abstract) context. By contrast, according to Conflation Theory, the elements of meaning traditionally identified as the source and target of the metaphor will only be separated out later in the child's development. Johnson conjectured that this influences the way in which these concepts are related in subsequent experience, and this may account for the degree of conventionality that VISION metaphors have. Even if they are not as productive or system-wide, it looks likely that senses other than vision may also be linked with mental processes at a very early stage. There is obviously important cultural input as well, and in part this may account for the huge bias towards VISION. Western society assigns vision such a privileged status that it has been described as "ocularcentric" (Jay 1993:4), and this is evident in all sorts of ways historically, including preoccupation with signs and symbols, belief in the authority of the written word and, in modern times, dependence on visual media such as TV and film. It must be the case that, as well as reflecting it, this perpetuates and intensifies the way we value and trust the visual over, for example, the auditory.

In her study of perception vocabulary, Sweetser points out that the link between vision and intellection is ancient, extending back as far as Proto-Indo-European. This is certainly the case, but investigations that I have carried out suggest that, in fact, the roots of some central vision vocabulary refer to both physical and mental vision as far back as they can be traced. The five most productive roots for my data are listed below, with their reflexes.

PIE ***weid-** (> PDE wise, wit, vision)
PIE ***ghel-** (> OE gl_aw > EME glew)
PIE ***sekw-** (> OE s_on > PDE see, sight etc)
PIE ***kap-**(> L capere > PDE perceive, conceive)
PIE ***sep-**(> L sapere > PDE sage, sapient)

Early examples of vocabulary, and evidence from cognates, seems to indicate that for three of these five roots it is not possible to say with any confidence that the physical sense came before the mental sense; in other words, one cannot find clear evidence of a metaphorical mapping from one concept to another, and this means that both concepts may always have been active for the root. I have identified a number of other roots, some with descendants in the data as well as others in the same semantic field that exhibit a similar connection with the physical and mental. Moreover, from a relatively cursory investigation of some non-Indo-European language families, it seems plausible that the same link is present, and this would strongly support a basic experiential motivation.

It seems to me that the connection between the senses and intelligence may not be a straightforward case of metaphor as it is traditionally understood, and I would contend that it requires further consideration. Any mapping from source to target (at least as these terms have most commonly been used) implies extension from one concept, the earlier meaning of a lexical item, to a second concept, which will develop as a later meaning of the same lexical item. If there are a number of instances of SENSES vocabulary for which the physical meaning does not precede the mental meaning developmentally or historically then this raises issues about the way in which metaphorical mappings can be described, and the differences between types of mappings.

ANIMAL

meaning	core concept	entry	part of speech	date
stupid	MAMMAL	**ape**	n	c1330-1741
stupid	MAMMAL	**sheepish**	aj	c1380-1692
stupid	MAMMAL	**mule**	n	c1470
stupid	INSECT	**hoddypeak**	n	1500-1589
stupid	BIRD	**daw pate/dawpate**	n	a1529-1562
stupid	MAMMAL	**ass-headed**	aj	1532+1609
stupid	FISH	**cod's head**	n	1566-1708
stupid	MAMMAL	**calvish**	aj	1570-1834
clever	MAMMAL	**shrewd**	aj	1589>
stupid	BIRD	**cuckoo**	n	1596>
stupid	MAMMAL	**long-eared**	aj	1605>
stupid	MAMMAL	**dunderwhelp**	n	1621+a1625
stupid	MAMMAL	**buffle**	n	1655+1710
clever	BIRD	**eagle-wit**	n	1665
stupid	BIRD	**dove**	n	1771
stupid	MAMMAL	**tup-headed**	aj	1816
clever	MAMMAL	**varment**	aj	1829>
stupid	MAMMAL	**bovine**	aj	1855+1879
stupid	FISH	**gubbins**	n	1916>
stupid	FISH	**like a stunned mullet**	aj	1953>

Animals, in the widest sense of the term, are one of the richest metaphorical sources in English and other languages. At every level of society, people are described as animals of all kinds: one can encounter *cows, dogs, sharks, worms, rats, weasels* and *lambs* in everyday experience, and there are few animals that cannot be related to humans in some meaningful way. The core category group ANIMAL accounts for 99 entries in total, making up something under 10% of the total data, and split in the following way:

concept	**entries**	**% of ANIMAL data**	**% of total data**
ANIMAL	99	100	9.20
MAMMAL	39	39.39	3.62
BIRD	36	36.36	3.35
INSECT[1]	14	14.14	1.30
FISH	9	9.09	0.84

Almost all of the data is used to signify stupidity – 92 entries compared with only seven signifying cleverness – and strikingly, of these seven, one is used in a derogatory way, and all of the others bar one are identified with sharpness or shrewdness as opposed to other kinds of cleverness. The exception is eagle-wit, and this has only one supporting quotation, dated 1665, in the OED. Clearly, then, in the rare cases where intelligence is associated with animals, the resulting terms tend towards a particular type of intelligence. Sharpness and shrewdness seem to indicate a worldly, practically applied cleverness, and perhaps also a certain lack of trustworthiness. There is an implication that, in terms of mental faculties, it is not natural for animals to be associated with humans, so that if they are it cannot be entirely positive. The Great Chain metaphor, discussed by Lakoff and Turner (1989:167ff), is important here; employing any animal metaphor for a person (or at least any derogatory one) can suggest that they are a 'lower' creature, in some way sub-human.

Perhaps more than any other within the data, the ANIMAL group illustrates the complexity that can be involved in a seemingly simple mapping. As Grady has pointed out (1997:222), the mapping is rooted in perceived similarity between animal and human, and this perception must be facilitated in some way. Although I would not argue that it is a conscious linear process, the basic principle of animal metaphorization can be broken down into several key elements, which combine to form an intuitive, gestalt-like source of description of people as animals.

1. This group includes snails, which are not insects in the technical sense but more correctly gastropods. I would contend that for most people these belong in the same working category and, for the purposes of simplification, are best seen as part of the same group.

At the basis of this are two general human tendencies in dealing with the world. The first of these is personification, ie the way in which humans ascribe, more or less deliberately, human qualities to non-human entities, presumably in order to best relate to them. Personification is obviously very common, and is evident in the way we deal with all sorts of entities. Some specific examples can be found in mappings identified in metaphor corpora: IDEAS AS PERSONS OR OTHER ANIMATE BEINGS (Barnden 1997), MACHINES ARE PEOPLE (Lakoff 1994) and THEORIES ARE PEOPLE (ibid.). In the case of animals specifically, this becomes anthropomorphism, and as anthropologists and archaeologists have pointed out, modern humans have been compulsive anthropomorphizers as far back as their history can be traced; there is evidence that in some circumstances this may even work as a survival mechanism (see, for example, the discussion of early hunting practice in Mithen 1998:190ff). Coupled with this tendency is a second process common to human conceptualisation, and this is our propensity to reduce entities to a single feature, selected on the basis of what appears most typical or distinctive. Lakoff & Turner refer to this as the "quintessential property", and give the examples of piety as quintessential to saints, filthiness as quintessential to pigs, and courage as quintessential to lions (Lakoff & Turner 1989:196). This is fundamental to a huge number of metaphors, and many explicit examples can be found in formulaic similes (of the form *as _ as a _*), where a single property is picked out and implied as the defining characteristic of an entity, very often an animal.

The result of the combination of these tendencies is that any animal can be widely understood to exemplify a particular human behaviour or characteristic. It seems natural that the direction of the mapping can then be reversed, so that the animal becomes a source of metaphors for humans. The 'similarity' on which any metaphor is based is not generally related to any scientific or factual reality about an animal's behaviour, since this is interpreted according to human expectations and intuitions; rather, the associations that particular animals acquire are influenced strongly by background and culture and tend to become part of the shared folk knowledge of a community. The availability of animals to be exploited as sources depends mainly on familiarity: for example, around a quarter of the mammal entries come from sheep, cattle and donkeys, all well established as farm animals in the UK, and with the exception of ape all of the others are woodland or similarly common animals. There is also a question of status, and since the target concept here is stupidity, all of the animals in the group either have a very specific and restricted purpose for humans, like the farm animals, or are low-value 'pests' as in varment (from *vermin*) and squirrel-headed.

An interesting feature of the data is that, although all the factors involved in this mapping seem to be based on basic human experience, the entries in this section are almost all dated to the sixteenth century or later. To a certain degree this must reflect the nature of written sources that survive from the OE and EME periods, on which my data relies. The majority of early medieval texts appear to have been fairly formal in register, and many dealt with biblical material, so it is perhaps unsurprising that metaphorical animal terms, which are likely to have been associated with slang and the spoken word, are not found until later. But cultural and societal influence must also be a factor. There was a well-established tradition of human-animal thought long before animal metaphors became conventionalized in language, and evidence of this can be found in classical works like Aesop's Fables and the bestiaries, both familiar and popular by the middle ages. But when the role of animals changed and interest in zoology increased, animals seem to have become more 'available' as source concepts.

Although the mapping between animals and intelligence has to some extent an experiential basis, this is of a very different order to that found in the SENSES data, since it results from the interaction of a number of basic cognitive mechanisms as well as important cultural factors.

DENSITY

meaning	core concept	entry	part of speech	date
stupid	WOOD/TREE	**stock**	n	1303>
stupid	GENERAL TERMS	**gross**	aj	1526-(1844)
stupid	WOOD/TREE	**log-headed**	aj	1571+1926>
stupid	FOOD	**groutnoll**	n	1578-1658
stupid	EARTH/TURF	**clod-poll/clod pole**	n	1601>
stupid	FOOD	**beef-witted**	aj	1606
stupid	EARTH/TURF	**turf**	n	1607
stupid	WOOD/TREE	**wattle-headed**	aj	1613+1866
stupid	EARTH/TURF	**muddish**	aj	1658+1829
stupid	WOOD/TREE	**a piece of wood**	n	1691

stupid	FOOD	**pudding-headed**		aj	1726-1867
stupid	MISC	**stunpoll**		aj	a1794>
stupid	WOOD/TREE	**nog-head**	n		c1800-1893
stupid	GENERAL TERMS	**dense**	aj		1822>
stupid	WOOD/TREE	**timber-head**		n	1849
stupid	WOOD/TREE	**off his chump**		aj	1877>
stupid	MISC	**ivory dome**		n	1923(>)
stupid	FOOD	**suet-headed**		aj	1937(>)
stupid	GENERAL TERMS	**thickie**	n		1968>
stupid	WOOD/TREE	**woodentop**		n	1983>

Of the three source concepts discussed here, DENSITY is certainly the one that has received the least attention. This is a surprising oversight, since it accounts for a good proportion of the data here, and my impression is that it is still highly productive in the way stupidity is conceptualised, indicated by the recent appearance of phrases like *thick as shit* and Scots *thick as mince*. All of the entries in this group signify stupidity, and a noticeable feature of the data is that there is no symmetrical concept to signify intelligence (in fact, items based on the source concept of loose texture, such as the dialect word fozy, also signify stupidity!). I have divided the entries into the following categories according to the substance from which they are derived.

concept	**entries**	**% of DENSITY data**	**% of total data**
DENSITY	93	100	8.64
WOOD/TREE	37	39.78	3.44
GENERAL TERMS	18	19.35	1.67
FOOD	18	19.35	1.67
EARTH/TURF	14	15.05	1.30
MISC SUBSTANCES	8	8.60	0.74

The motivation for this group is image-based, and works almost like a narrative. The idea is presumably that if something is dense in its physical texture, it will be difficult to penetrate, so if a person's mind is dense, ideas and knowledge cannot easily get in or through. This has a number of entailments, dependent on certain other metaphors fundamental to the way the mind is conceptualised. For the mind to have any sort of texture, it must be a physical, bounded entity, and this is a common and well documented mapping: ATT-Meta lists MIND AS PHYSICAL ENTITY, and Lakoff's Conceptual Metaphor Homepage includes THE MIND IS A BODY and THE MIND IS A MACHINE, both specific examples of this. For things to get 'through' the mind's boundary and 'inside' it, a container schema must be closely aligned with the mapping. This fits in with other core category groups within the data, including CONTAINER itself, as well as the entries relating to grasping, which I referred to above within the SENSES group – a basic way of accounting for grasp is roughly as a blend of TOUCH and CONTAINER. A common mapping related to the container metaphor is IMPORTANT IS CENTRAL (Grady 1997:284), and this seems relevant as well.

The interesting thing about the group is that the source concepts from which individual entries derive are surprisingly specific, and there are a very limited number of these. As the above figures indicate, less than 20% of the group is made up of general terms, such as thick, crass, dense etc, and within the rest of the group almost all the entries are connected with one of the three groups WOOD/TREE, FOOD and EARTH/TURF. These are all are commonplace, mundane substances, that have little value in their crude state (the FOOD group contains entries connected with basic ingredients and uniform consistency, like pudding-headed and beef-brained, rather than more complex foodstuffs). As well as this, it seems to me that the idea of cognitive 'cohesion' is helpful here. I have judged these entries to belong to a single core concept group because they seem to me to have a basic property in common, but I acknowledge that the source concepts involved in this group are not suitable to express lack of intelligence only because they are dense substances. Other properties, such as the fact that they are raw, uncrafted materials like wood, or formless masses like earth, must also be relevant, and perhaps the combination of properties make them more cognitively 'convincing', especially since these are not selected as a result of conscious reasoning about motivation.

One question that presents itself here is why certain dense substances are less successful as sources even though they would seem to be equally as available and suitable as those that do appear. For example, stone would seem to be ideal, since it is particularly dense in composition whilst being similar to the source substances in value and commonness. However, there are only two entries derived from stone

(listed within the MISC section): stone, which has a single supporting quotation in 1598, and stunpoll, which is cited in 1794 and continues into current usage. The *OED* suggests uncertainly that stunpoll is derived from a variant of *stone* in compound with *poll*, head, but it should be noted that folk etymology would be likely to associate this with the verb *stun*, especially since there are other items in the data like stupid itself which can be traced back to Latin *stupere* 'to hit, stun'. This must be a factor in its continued use. I would speculate that there may be various reasons for the lack of any other stone entries. It may be simply too hard – although substances like wood and lumps of earth are dense, they can be penetrated with effort, whereas stone is a completely different texture (as is bone, which yields only three entries). Correspondingly, there is a difference between being able to comprehend something with difficulty (ie get it 'into one's brain') and being wholly incapable of this. Aside from this, and perhaps more convincingly, there may be an issue about other properties metaphorically associated with any entity. Stone is commonly used as a source concept for steadiness and constancy, as when someone is described as a *rock* or *brick*, and equally it can be used to connote cruelty and indifference, as in a *heart of stone* or a *stony expression*. It may be that this precludes its selection for other target concepts. To an extent, the characteristics that come to be associated with particular entities, and conversely the selection of specific entities over others to connote these characteristics, must be arbitrary, even though the general mapping may be clearly motivated. This demonstrates the way in which the shared associations of a community are crucially important, and must be taken into account in the analysis of any metaphorical mapping.

Conclusion

In the light of the data I have studied, it seems to me that 'metaphor' is most practically useful if employed as a broad, inclusive term. Steen argues that a conceptual definition of metaphor has implications for the way metaphors in discourse are regarded, asserting that "Conceptual metaphors may emerge as linguistic metaphors, similes, analogies, extended nonliteral comparisons and allegories, to name only the most obvious possibilities. Other divisions include personification, synaesthesia, and zeugma, while there are also the related categories of proverbs, sayings, idioms and symbols" (Steen 2002:21). I believe that this approach has value, and I would argue that at its simplest level 'metaphor' can be used even more broadly, to cover metonymy, synecdoche and simile as well. This is not to say that metaphor is necessarily the most basic or conceptually important process, or that these other terms are not useful, but it is a fact that metaphorical mappings can result from quite varied mechanisms and can therefore be diverse in nature. It is crucial to recognise that, if they are to be collected together, the term used to label the resulting group must be able to accommodate this diversity. If this is the case, it does not seem unreasonable to subsume other mappings that rely on similar mental processes under the same collective name, if only for the sake of economy. Moreover, it is my impression that this is the way that 'metaphor' is often used in practice, certainly by non-linguists (as represented by the dictionary definitions quoted above), but also by metaphorists themselves. This can result from difficulties in determining precisely which kind of mapping is involved in particular cases (see Feyaerts 1999:319 for comments on the connection between SEEING and KNOWING), and from blending, frequently of metaphor and metonymy (cf. Goosens' (1990) use of the term 'metaphtonomy'). Even aside from these complications in classifying particular linguistic items, though, 'metaphor' tends to be used as the generic for a particular group of phenomena.

In employing a more practically useful, less restrictive definition of metaphor, efforts can be diverted away from issues of classification which are, by nature, unlikely to be resolved. Instead, there can be sharper focus on the analysis and deconstruction of metaphor and its relationship with cognition, which promises to be a much more productive and significant area for investigation.

References

BARNDEN, J. 1997 *ATT-Meta Project Databank: Examples of Usage of Metaphors of Mind.* http://www.cs.bham.ac.uk/~jab/ATT-Meta/Databank/index.html

FEYAERTS, K. 1999. Metonymic hierarchies: the conceptualization of stupidity in German idiomatic expressions. In Panther, K-U. & Radden, G. (ed.) *Metonymy in language and thought. Amsterdam: John Benjamins, pp309-332.*

GOOSSENS, L. 1990. Metaphtonymy: the interaction of metaphor and metonymy in expressions for linguistic action. In *Cognitive Linguistics* 1(3), pp323-340.

GRADY, J.E. 1997. *Foundations of meaning : primary metaphors and primary scenes.* PhD Thesis, University of California, Berkeley.

JOHNSON, C. 1999b. Metaphor vs. conflation in the acquisition of polysemy: the case of *see. In Hiraga, M.K., Sinha, C. & Wilcox, S. (eds.) Cultural, psychological and typological issues in Cognitive Linguistics: selected papers of the bi-annual ICLA meeting in Albuquerque, July 1995. Amsterdam: John Benjamins, pp155-169.*

KAY, C.J. 2000. Metaphors we lived by: Pathways between Old and Modern English. In Roberts, R. & Nelson, J. (eds.) *Essays on Anglo-Saxon and related themes in memory of Lynne Grundy. London: King's College London, Centre for Late Antique & Medieval Studies, pp273-285.*

LAKOFF, G. 1994. *Conceptual Metaphor Homepage.*
http://cogsci.berkeley.edu/MetaphorHome.html

LAKOFF, G. & JOHNSON, M. 1999. *Philosophy in the flesh: The embodied mind and its challenge to Western thought.*. New York: Basic Books.

LAKOFF, G. & TURNER, M. 1989. *More than cool reason: A field guide to poetic metaphor.* Chicago: University of Chicago Press.

MITHEN, S. 1998. *The prehistory of the mind: a search for the origins of art, religion and science.* London: Phoenix.

SIMPSON, J. & WEINER, E. (eds.) 1989. *The Oxford English Dictionary* (2nd ed.). Oxford: Oxford University Press.

STEEN, G. 2002. Towards a procedure for metaphor identification. In special issue of *Language and Literature*, *'Metaphor Identification'*, 11:1, pp. 17-33.

SWEETSER, E.E. 1990. *From etymology to pragmatics: metaphorical and cultural aspects of semantic structure.* Cambridge: Cambridge University Press.

Historical Thesaurus of English homepage:
http://www.arts.gla.ac.uk/SESLL/EngLang/thesaur/homepage.htm

NLP model and tools for detecting and interpreting metaphors in domain−specific corpora

P. BEUST, S. FERRARI, V. PERLERIN
GREYC, Computer Science Laboratory, University of Caen − bd Maréchal Juin
F14032 Caen Cedex − France

Abstract

The aim of this paper is to present how a user−centred lexical representation model, based on the theory of Interpretative Semantics, can be used for detecting and interpreting metaphors in domain specific corpora. We present here several tools useful for such tasks and discussing the results of an experiment.

Introduction

In this paper, we present NLP (Natural Language Processing) project addressing the interpretation process. This project, called "IsoMeta[1]", focuses on computer−assisted metaphor interpretation following a user−centred point of view. We propose a model for lexical representation as well as tools for validation on corpora.

In the first section, we give an overview of some previous approaches related to metaphor detection and interpretation in order to highlight the main concepts we deal with. We also introduce the theoretical background for knowledge representation and text interpretation sustaining our approach.
In the second section, we argue for user−centred lexical representations and we present our model for this purpose (called *Anadia*) as well as practical examples. This model enables automatic computing of customized help for interpretation by means of the isotopy concept. We detail how to produce such help when dealing with conventional metaphors.
In the third section, we present some of the tools implementing our main propositions. AnadiaBuilder is a user−friendly interface to build structured lexical representations. Complementary tools have been developed for corpus analysis, producing graphical representations for easy browsing through the results and customized help for interpretation.
In the last section, we present the results of an experiment on a domain−specific corpus. We study examples of a specific conventional metaphor: the stock market domain expressed with meteorological terms.

Finally, we discuss how to carry out an evaluation of our work. We also propose other applications of our model and tools. We conclude by pointing the main directions for further developments and the next steps for the "IsoMeta" project.

1 Framework

1.1 Metaphors in NLP

It is generally agreed that a metaphor involves two concepts: a source concept, related to the words used metaphorically, also called the vehicle of the metaphor, and a target concept, which is what the metaphor is used for and tries to describe, also called the tenor of the metaphor. If we consider the following example, first proposed by Wilks (1978), and still studied by Fass (1997):

(1) "My car drinks gasoline",

the source of the metaphor is *the action of drinking*, and the target may be described as *the use of gasoline by a car*.
The different NLP approaches for metaphor interpretation mainly depend on how the relation between the source and the target is viewed: as an analogy, as a novelty, or as an anomaly. In (Gentner, 1983; Falkenhainer et al., 1989), this relation is mostly viewed as an analogy. Thus, interpreting a metaphor requires deeply structured knowledge representations in order to trace back and describe the analogy between concepts. In (Indurkhya, 1992; Gineste et al, 1997), the relation between the source and the target is viewed as a novelty: it is not a pre−existing similarity but one created by the existence of the metaphor. Thus, interpreting it requires the dynamic selection and transfer of knowledge from the source domain to the target domain.

[1] "IsoMeta" stands for Isotopy and Metaphor.

Metaphor may also be viewed as a semantic anomaly. In example (1), there is an anomaly if one considers that "*drinking*" does not normally apply to physical objects such as cars. As shown by Martin (1992), metaphors are not always anomalies, and anomalies are not always metaphors. For instance, in:

(2) "McEnroe killed Connors" (ibid),

there is no anomaly, nonetheless "killed" may be viewed as metaphoric. Only contextual information can help for disambiguating the whole sentence. Fass (1997) proposes a method for discriminating semantic relations, which makes a clear distinction between metaphors and anomalies. This method makes it possible for multiple interpretations to coexist, as in example (2).

It is not necessary to focus on the relation between the source and the target to interpret metaphors. Kintsch (2000) shows how the meaning of a metaphor can be interpreted and represented by a multi−dimensional vector, exactly like other meanings in the Latent Semantic Analysis approach.

We also consider that metaphors require the same interpretation process as other meanings. We do not focus on the relation between the source and the target either. But in our approach, we use a symbolic representation in order to provide a novice user with easily understandable tools.

Lakoff and Johnson (1980) introduced the notion of conventional conceptual metaphor, based on the observation that, for some semantic domains, multiple terms from a common source domain may be used to describe metaphorically multiple corresponding concepts from a common target domain. In (Ferrari, 1997), such conventional metaphors are studied in the scope of domain specific corpora. For instance, he observed that stock market events are often described by meteorological terms in newspaper articles related to economics.

In our work, we look at conventional metaphors in order to use the pre−existent knowledge that the target domain may be partly structured as the source domain. We focus on the previous example, which we call "*economics is meteorology*".

Using limited and user−centred resources, we try to track down the analogy and the novelty points of view. In the next section, we present the linguistic basis of our approach.

1.2 Knowledge representation and text interpretation

The lexical representation and the analysis process we use are mainly inspired by continental structural linguistics (Greimas, 1966; Pottier, 1987) and especially by the linguistic theory developed by F. Rastier (1987): Interpretative Semantics. In this theory, the interpretation is considered as a description of semantic units located both in a linguistic unit (corpus, text, sentence...) and a situation. Interpretation involves an interpreter, along with his knowledge, his goals and his social relation[2] to these given linguistic units. Thus, the meaning of a word, for instance, is not a definition of this word, as could be found in a dictionary, but rather an explanation of its role in a given linguistic unit.

A lexical content is described in terms of meaning components, themselves described in terms of semantic features called *semes*. For example, the lexical item "depression" can be related to a ’meteorological phenomenon’ or a ’mental state’, and the meaning component ’meteorological phenomenon’ can be represented with the following *semes*: /area/, /low pressure/, /bad weather/? Such a description is called a componential representation.

Semes depend both on the user and on the task. They are potential meaning features, relevant only in specific contexts. The notion of *isotopy*, introduced by Greimas (1966), characterizes these contexts. An isotopy is the recurrence of one *seme* in a linguistic unit. For instance, in this paper, one may notice at least two main isotopies related to ’computer science’ and ’linguistics’, supported by many different lexical items.

In our work, we focus on lexical items from two domains, meteorology and stock market, in order to describe the underlying conventional metaphor. In the next section, we present Anadia, the model we have previously developed for such lexical representations, and show how to use it for metaphor processing.

2 A model for lexical representation

2.1 Main principles

The main principles of our model have been described in details by Beust (1998) and Nicolle et al. (2002). Anadia is a model of lexical categorization based on both componential and differential representation. The differential paradigm states that a lexical content can be described by opposing it to others through structural relations, following the notion of "linguistic value" proposed by

[2] We are talking about the relation to linguistic units through social role. For instance, a juridical text is differently interpreted by a lawyer and by common people.

Saussure (1915). The *Anadia* model allows a user to produce descriptions of meaning components by the way of *semes*, which are the componential part of the representation. Rather than classical componential representations, *semes* are represented by a set of opposite features. This is the basis of the differential part of the representation. For example, "*depression*" can be described as the combination of the *semes* [Zone] and [Pressure] respectively corresponding to the opposite features "area vs. line" and "low vs. high". The activated features for "*depression*" are area and low. These *semes* also allow a semantic representation of the lexical item "*anticyclone*" described by the activated features area and high.
Lexical items representations are therefore made from the combination of *semes*. In this way, our model allows its user to build tables where lexical items can be described in terms of differences and common points, as shown in Figure 1.

Pressure zone	Zone	Pressure
anticyclone	area	high
tropical wave, easterly wave	line	low
depression, cyclone	area	low
	line	high

Figure 1. Example of an Anadia table describing some pressure zones[3].

In Figure 1, the combination of the *semes* [Zone] and [Pressure] gives rise to four table rows in which lexical items can take place. When there are several lexical items in the same row, it implies that their semantic representations are not considered as different in this table (in another one, they could be differentiated). It is the case for "tropical wave" and "easterly wave" in the example. A row can stay empty if we do not know any lexical item corresponding to a certain combination of features. It is the case for the combination of 'line' and 'high' in the example. A row can also be filled in later if we find a corresponding lexical item (for instance, by the way of a corpus study).

Several tables can be used to describe a specific semantic domain. In such a set of tables, a table can be linked to a row in another table by a subcategorization relation (Figure 2).

Domain objects	**Role**
stocks, currency (…)	playing a part
charts, ratio, stock indices, curves (…)	studying, analysing

Stock Indices	**Geographical Zone**
CAC, CAC40	France
Dow Jones, Nasdaq	USA
Nikkei	Japan
Dax	Germany
Footsie	UK

Figure 2. Extract from a set of tables for the stock market domain.
The second row of the Domain objects table is linked to the Stock indices table by a relation of subcategorization.

For many reasons (choice of *semes*, content of rows, subcategorization relations) tables represent the points of view of the user for a given task. Anadia is a user−centred model and the lexical representations built with the model are not supposed to be either universal or exhaustive. Tables can be modified and updated at any time, depending on the results obtained from the analysis process.

Anadia tables allow proposing an analysis process based on the concept of isotopy. As shown by Tanguy (1997), isotopy can be seen as an easy and understandable way of expressing themes in linguistic units. Therefore, the interpretation process consists in finding isotopies in linguistic units.

> *(3) During the three days immediately proceeding **depression** formation, anomalous moisture transforms from a pattern associated with a **tropical wave** transversing the open Atlantic Ocean ... (http://ams.confex.com/ams/25HURR/25HURR/abstracts/35268.htm)*

In example (3), using the representation of Figure 1, we notice that "tropical wave" and "depression" are described with the same *semes*: [Zone] and [Pressure]. These two recurring *semes* involve two isotopies

[3] The examples have been translated for this paper.

that contribute to the meaning to the sentence. The recurring features also show that the sentence deals with pressure zones of different type : one corresponding to a ’line’ of ’low’ pressure and one to a ’area’ of ’’low’ pressure.

2.2 Using the model for metaphor processing

The *Anadia* model was not originally designed for metaphor processing. The latter is just a specific task for which the model can be used. In order to study how the model can effectively be applied to metaphor processing, and what adjustments are to be made, we focus on the specific conventional metaphor: *"economics is meteorology"*.

The model enables us to represent our lexical knowledge concerning the source and the target domains involved in this specific metaphor. Let us work on the assumption that one set of tables, set S, describes the lexical items of the source domain, *meteorology*, and a second one, set T, is dedicated to the target domain, *stock market*.

At this point, the Anadia model enables us to use a single lexical item in multiple sets of tables. For instance, it is possible to represent "*barometer"* both in set S and in set T. In set S because it is a common term of meteorology, and in set T because we have noticed in newspaper articles that it is sometimes used in phrases such as "*stock market barometer*", suggesting some economical tools for measures or predictions.

This possibility becomes a problem when dealing with metaphors. If we want to use the model to detect the metaphorical use of *"barometer"* in phrases such as *"stock market barometer"*, we must not represent it in set T. Moreover, lexical items of set T must not be formed with words that can be considered as lexical items of set S. This is a first adjustment, or constraint, added to the Anadia original model: when building sets of tables for metaphor processing, it is necessary not to use words from a source domain in a set of tables for a target domain.

Following this rule, *"barometer"* is now banished from the lexical items of set T. The reason for this is that when computing isotopies, the source *semes* are required to spot a metaphorical use. If *"barometer"* were in the two sets, S and T, its metaphorical use in *"stock market barometer"* would be ignored because an isotopy of words from set T would only hide the existence of *semes* from the source domain. It is important to notice that such a representation must not be considered as "wrong" and would not lead to misinterpretation. It would simply reflect the conventional aspect of the metaphor, which itself would be part of the knowledge of the user who would include "*barometer"* in the lexicon related to *"stock market"*.

Assuming that S and T are now built according to that constraint, let us see how it is possible to spot a metaphor, and to what extent the lexical representation can produce guidance for its interpretation. The whole point is to detect an isotopy involving words from both the source and the target domain. On the one hand, with the Anadia model, isotopies are based on *semes* shared by lexical items involved in a single linguistic unit. On the other hand, previous works on conceptual metaphors have shown the existence of underlying structure analogies between the source and the target domains. It then stands to reason that the solution is to use some *semes* which are shared by lexical items from the two sets of tables, and which represent the structure analogy between the two domains. For example, if we use the *seme* [Role = studying, analysing vs playing a part] to describe "*barometer*" from the meteorology domain and "*stock exchange*" from the stock market domain, it then becomes possible to spot and produce guidance for interpreting the following metaphor :

(4) a− "the Dow Jones is a stock exchange barometer".

The *seme* [Role] is here shared by two lexical items: *"barometer"* from the source domain and *"Dow Jones"* from the target domain. The fact that the lexical items involved belong to different domains is characteristic of a metaphorical use. The shared *seme*, creating an isotopy, is a first step for guiding the interpretation process. We shall discuss these points further in the following sections.

At the moment, we can consider the use of shared *semes* as a second adjustment or constraint added to the model when processing metaphors. If sets S and T are built according to the two constraints presented in this section, it is not only possible to spot metaphors involving the lexical items initially used to organize the two sets, but also to process some of their extensions. Actually, when building the set of tables concerning *meteorology*, the user will probably consider lexical items such as "*thermometer*", "*mercury*", and propose to use the same *seme* [Role] to describe them. It will then be possible to process the following examples:

(4) b− "the Dow Jones is a stock exchange thermometer"

(4) c− "the Dow Jones is the New York stock exchange mercury"

even though the sets of tables were not originally designed for these specific metaphors.

The next section presents tools developed in order to validate our model on corpus.

3 Tools

The tools we created for our experiments are freely available for research purposes. They have been implemented with platform−independent languages (Java, XML and XSL). They can be used for different kinds of tasks including figurative language analysis (as shown in this paper) or for instance document retrieval (as shown in (Perlerin, 2001)).

3.1 AnadiaBuilder: a tool for building Anadia lexical representations

AnadiaBuilder is software enabling to build lexical representations following the Anadia model (Nicolle et Al., 2002). The created data is stored in XML format. Via a user−friendly graphical interface, the user can build sets of tables according to the current task. The interface contains five main interactive panels:

(A) The first one enables the user to create the *semes* he finds relevant for the representation. The user chooses the related sets of opposed features and an explicit name for each *seme*.
(B) The second one makes it possible to create tables made from the combination of semes (Figure 3). The user chooses the semes and the machine computes the combinations and automatically builds the table. The user fills in the cells (on the left−hand part of the table) corresponding to a given set of features from different semes with relevant lexical items.
(C) The third one displays a graphical representation of a table (called "topique" in French) showing the differences and the semantic proximity between lexical items by means of annotated links (Figure 3).
(D) The fourth one creates the relations between tables. It also makes it possible to see the whole set of tables through a schematic representation where only table names are displayed (Figure 4). In this panel, the user can allocate a colour to each table, which is useful for further corpus analysis.
(E) The last one is linked to the MAHTLEX lexical database, developed at the University of Toulouse[4]. For each lexical item, the computer proposes a set of inflections or enables the user to build the corresponding set of inflections by himself. Inflections will be used to match occurrences of lexical items in texts.

At step (B), when building a table, if the user estimates that he can fill in several cells with the same lexical item, he must correct his proposals. This fact can happen because of two reasons. The chosen *semes* are not mutually exclusive, or the features of at least one *seme* are not mutually exclusive. The building contraints of the Anadia model are discussed by Beust (1998). Perlerin et Beust (2002) have undertaken an experiment with novice users. The results have shown that building a set of tables following the Anadia constraints is accessible to novice users. Such results may have to be moderated when dealing with a linguistic phenomenon such as metaphor.

[4]http://www.irit.fr/ACTIVITES/EQ_IHMPT/ress_ling.v1/accueil01.php

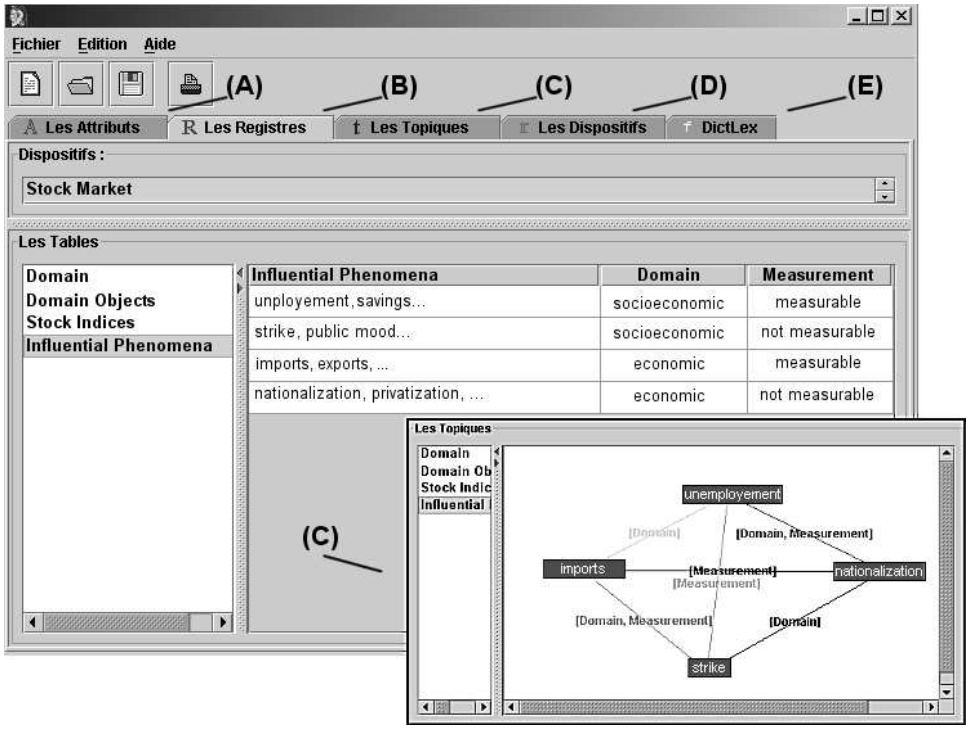


Figure 3. AnadiaBuilder: tables building panel and corresponding "topique" from the "topique" panel (extract of the screenshot).

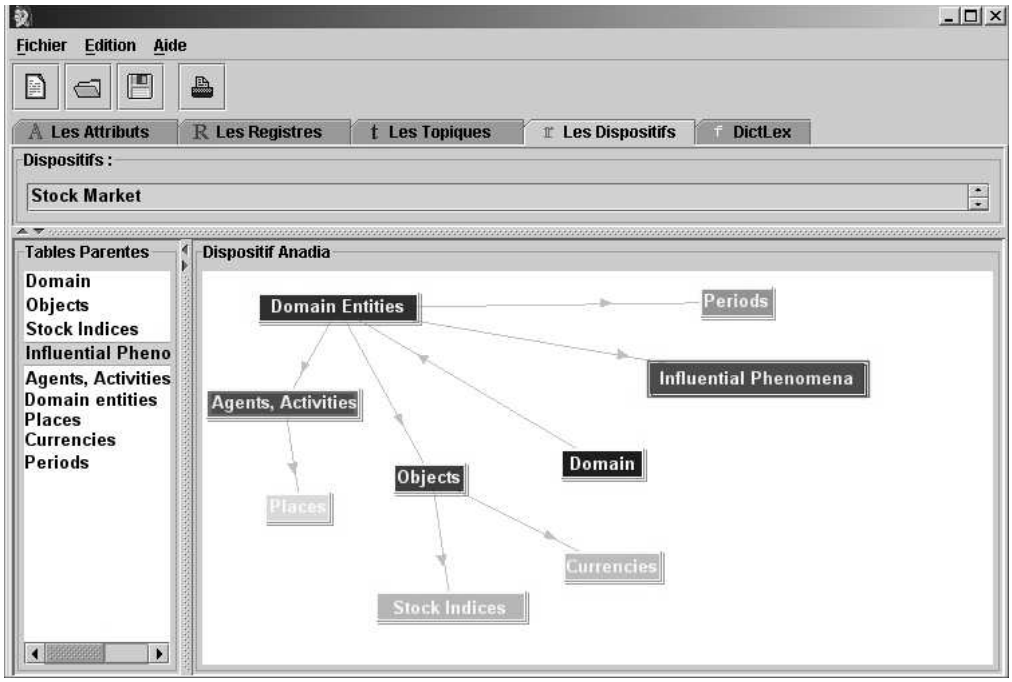


Figure 4. AnadiaBuilder: set of tables representation related to the stock market domain.

Each set of *semes*, each set of tables or inflections dictionary can be saved independently and reused in different experiments. In particular, the sets of tables can be used for corpus analysis. Results are then produced as an annotated version of the corpus. Several tools help us to browse through the resulting corpus, mainly by the use of colours and charts.

3.2. Corpus analysis tools

During the automatic part of the corpus analysis, all the possible occurrences of lexical items from the sets of tables are located in the texts. A first tool builds a graphical representation of each text in the corpus[5], as shown in Figure 5. For one text, each table is represented by one bar inheriting its colour. Each bar is proportional to the number of matched lexical items from the table. In our experiment on the metaphor "*economics is meteorology*", the purpose of this graphical representation is to provide the user with a quick way to track down articles where the source domain is evoked. A single HTML page contains all the charts along with hyperlinks to the related texts (Figure 5).

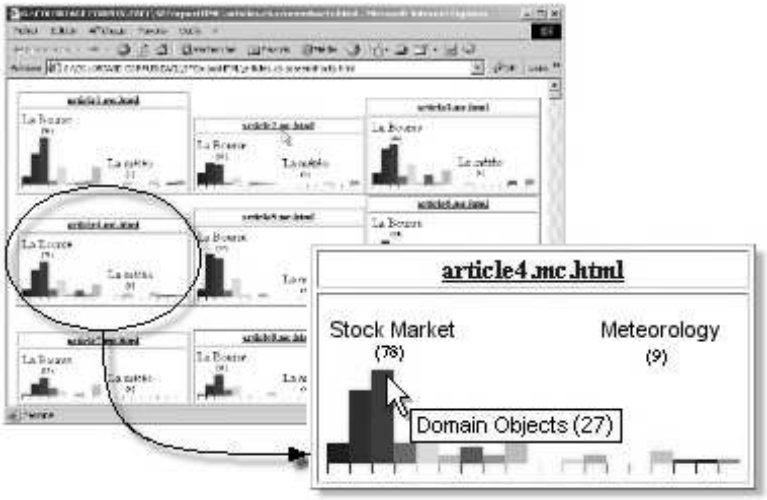


Figure 5. Graphical representations of the outputs:
moving the mouse over a bar shows the corresponding table name
and matches the number of lexical items.

A second tool transforms the XML version of each text into an HTML version, as shown in Figure 6. In the HTML version, the matched lexical items are in the same colour as the corresponding table. This provides the user with an easy means to find the precise location of the lexical items he is interested in.

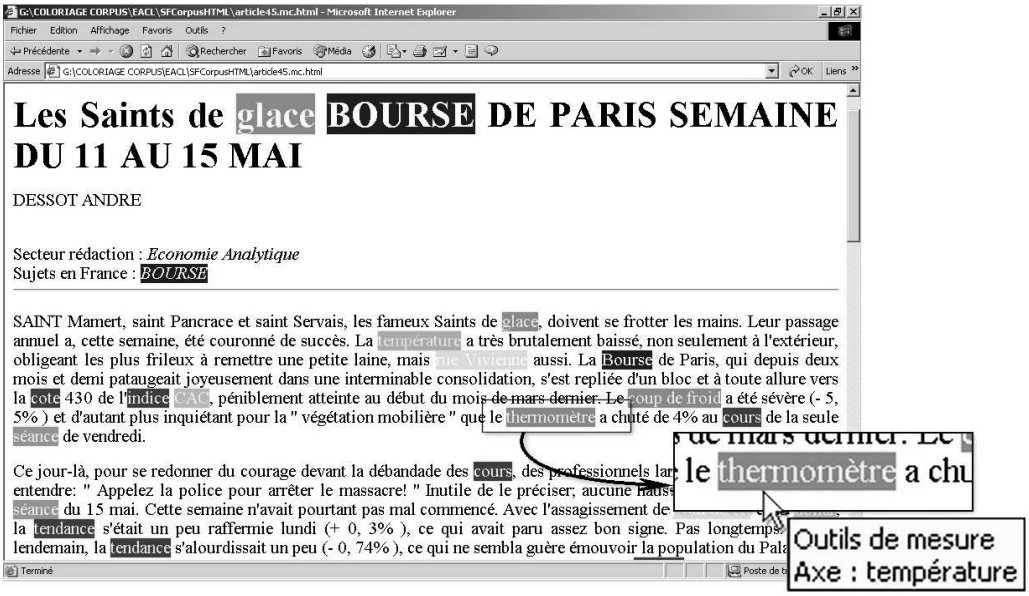


Figure 6. A coloured article.
Moving the mouse over a coloured lexical item shows the corresponding
table name and the corresponding set of *semes*/features.

The next session presents some results of an experiment realised on a journalistic corpus.

[5] In our experiments, the article appeared to be a relevant unit to build the charts. The level of this linguitic unit can be changed.

4 First results

Our work has been validated through a corpus experiment. The corpus is constituted of about 600 articles from the French newspaper "Le Monde", addressing economics and stock market (around 450,000 words) between 1987 and 1989. This corpus, already studied by Ferrari (1997), contains numerous examples of the conventional metaphor "economics is meteorology". It also contains lexical items from the meteorology domain that are not used in a figurative way.

For our experiment, the sets of tables have been designed with nine shared *semes*. These *semes* reflect our own view of the conceptual metaphor. Specialists of any of the two domains would probably have designed the sets of tables in a much different way. Our point of view reflects our knowledge of the underlying analogy between the two domains. In the following, we discuss two different examples in order to show how the analogy and novelty points of view can be retrieved with our proposals.

(5) *Le Dow Jones par exemple, le thermomètre de la Bourse de New York, qui avait chuté de 508 points ?*[6] − Article n°126 − Paragraph 1

In example (5), three lexical items from the sets of tables were matched (therefore coloured) by the analysis process. "Dow Jones" appears in the "Stock Indices" table of the stock market domain (see Figure 9). "thermomètre" (*thermometer*) appears in the "Measuring Instruments" table of the meteorology domain (see Figure 8).

Domain objects	Role
sky, pressure, temperatures …	playing a part
degree, bar …	studying, analysing

Measuring Instruments	Axis
anenometer	wind
pluviometer	rainfall
thermometer	temperature
mercury, barometer	pressure

Figure 8. Extract of the meteorology *Anadia* set of tables.

Domain objects	Role
stocks, currency (…)	playing a part
charts, ratio, stock indices, curves (…)	studying, analysing

Stock Indices	Geographical Zone
CAC, CAC40	France
Dow Jones, Nasdaq	USA
Nikkei	Japan
Dax	Germany
Footsie	UK

Figure 9. Extract of the stock market *Anadia* set of tables.

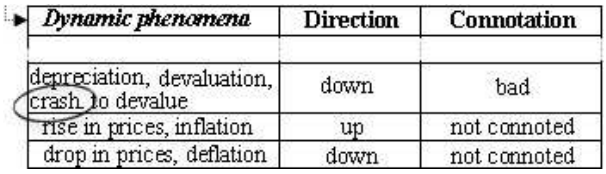
Following these representations of the two domains, an isotopy involves the shared inherited *seme* [Role] and the value 'studying, analysing' can be found thanks to the first two coloured lexical items. One can then conclude in favour of a metaphorical use and propose the following interpretation: "thermomètre" (*thermometer*) is used in the same way as "graphics", "ratio"... i.e. to suggest an object for analysis and study in the stock market domain. The lexical item could be replaced (more or less efficiently) by others from the "Measuring Instruments" table.

(6) *Ce était dû (?) à la chute vertigineuse et incontrôlée du , signe que la affecte dorénavant les* .[7] − Article n°153 − Paragraph 3

In example (6), the lexical items "krach" (*crash*) and "tempête" (*storm*) appear in the following tables (Figure 10 and Figure 11).

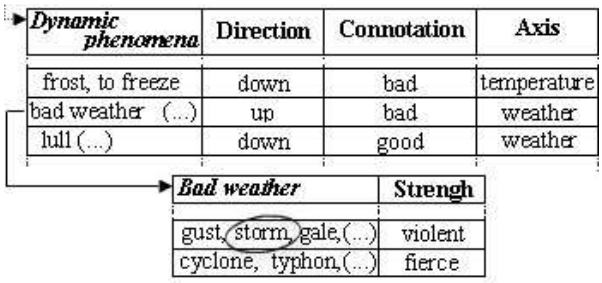
[6] Literal translation: *The **Dow Jones**, for instance, the **thermometer** of Wall Street, which had fallen 508 points ?*

[7] Literal translation: *This **crash** was due (...) to the vertiginous and uncontrolled fall of the **dollar**, sign that the **storm** will henceforth affect the financial markets.*



Dynamic phenomena	Direction	Connotation
depreciation, devaluation, crash, to devalue	down	bad
rise in prices, inflation	up	not connoted
drop in prices, deflation	down	not connoted

Figure 10. Extract from the stock market *Anadia* set of tables (the table has been truncated).



Dynamic phenomena	Direction	Connotation	Axis
frost, to freeze	down	bad	temperature
bad weather (...)	up	bad	weather
lull (...)	down	good	weather

Bad weather	Strengh
gust, storm, gale,(...)	violent
cyclone, typhon,(...)	fierce

Figure 11. Extract from the meteorology *Anadia* set of tables (the tables have been truncated).

The isotopy found in this sentence (example 6) is based on two different *semes*. The first *seme* involved is [Connotation] (inherited for "storm") with the same activated value ’bad’. The second one is [Direction] with two different activated values: ’down’ for "krach" and ’up’ for "tempête". Example (3) makes it possible to conclude in favour of a metaphorical use. First, due to the activated values, the *seme* [Direction] is less relevant than the other one, [Connotation]. Moreover the *seme* [Axis] is exclusively used in the meteorological domain and is not involved in any isotopy. We propose therefore to consider it as "irrelevant" in the context. The *seme* [Strength] does not take part in an isotopy either; but, unlike [Axis], it can be shared between several lexical domains. It seems to us that we can therefore consider it as relevant in this context. This illustrates how novelty is dealt with in our approach. Finally, we propose the following help for interpretation: "tempête" (*storm*) is used to evoke a not only bad but also violent dynamic phenomenon in the stock market domain.

Numerous examples of sentences where the sets of tables enable to conclude in favour of metaphorical uses have been discovered in the corpus thanks to our tools. The two sets of tables have been modified several times depending on the results obtained from the analysis process. Those results are the fisrt step of the "IsoMeta" project validating our approach and our tools.

Conclusion and further works

This paper has presented a user−centred lexical representation model and its use to produce help for metaphor interpretation. There is no need to be an expert in a given domain to describe it by means of this user−centred model. Nevertheless, metaphor interpretation is a linguistic task. Thus, a description for a study on a conceptual metaphor, such as the one we have presented in this paper, requires a certain familiarity with linguistic sciences. The user must indeed be able to describe how he appreciates the analogy between the source domain and the target domain by the use of shared *semes*.
Though we have presented the use of the Anadia model for a very specific task, we have already argued for its use in many applications, such as domain−specific corpus browsing or document retrieval, as shown in (Nicolle et al. 2002). We hope the same applies to the tools developed for the "IsoMeta" project. An experiment on domain−specific corpus has validated our method. Actually, producing customized help for metaphor interpretation appears to be possible. However, this result must be evaluated, both quantitatively and qualitatively. Nevertheless, such an evaluation is not easy to carry out. On the one hand, the user−centred aspect of the model implies that the evaluation process should be user−centred too. On the other hand, this evaluation requires an annotated corpus. Such a reference corpus does not exist yet and seems difficult to produce.
In order to start the evaluation, our further works will concern other examples of conceptual metaphors, as well as other domain−specific corpora for their study and the automatic processing of isotopies. We also plan to use our model for metaphor and paraphrase in automatic text generation.

References

Beust P. (1998). *Contribution à un modèle interactionniste du sens.* Computer Sciences PhD Thesis of the University of Caen, France.

Falkenhainer B., Forbus K.D. and Gentner D. (1989). *The Structure−Mapping Engine : Algorithm and Examples*. Artificial Intelligence, 41/1, pp.1−63.

Fass, D. (1997). *Processing metaphor and metonymy*. Greenwich, Connecticut: Ablex Publishing Corporation.

Ferrari, S. (1997). *Méthode et outils informatiques pour le traitement des métaphores dans les documents écrits.* Computer Sciences PhD Thesis of the University of Paris XI, France.

Gentner D. (1983). Structure−Mapping: A Theoretical Framework for Analogy. Cognitive Science, 7, pp. 155−170.

Gineste, M.−D., Indurkhya, B. and Scart−Lhomme, V. (1997). *Mental representations in understanding metaphors*. Technical report, 97/2, Groupe Cognition Humaine, LIMSI−CNRS, Orsay.

Greimas A.J. (1966). *Sémantique structurale*. Paris: Larousse.

Indurkhya, B. (1992). *Metaphor and Cognition*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kintsch, W. (2000). *Metaphor comprehension: A computational theory*. Psychonomic Bulletin & Review, pp. 257−266.

Lakoff G. and Johnshon M. (1980). *Metaphors we live by*. University of Chicago Press, Chicago, U.S.A.

Martin J.H. (1992). *Computer Understanding of Conventional Metaphoric Language.* Cognitive Science (16), pp.233−270.

Nicolle A., Beust P. and Perlerin V. (2002). *Un analogue de la mémoire pour un agent logiciel interactif.* In Cognito, 21, pp. 37−66.

Perlerin V., (2001) *La recherche documentaire, une activité langagière*. In proceedings of TALN2001, Tours.

Perlerin V. and Beust P. (2002) *Pour une instrumentation informatique du sens.* Proceedings of the CNRS/ARCO Summer School in Tatihou, to be published.

Pottier B. (1987). *Théories et analyse en linguistique.* Hachette: Paris , p. 224.

Rastier F. (1987). *Sémantique interprétative*. Presses Universitaires de France : Paris.

Saussure F. de (1915). *Cours de Linguistique Générale*. Mauro−Payot: Paris (1986).

Tanguy L. (1997). *Computer−Aided Language Processing: Using Interpretation to Redefine Man−Machines Relations*. Proceedings of the 2nd International on Cognitive Technology (CT'97), Humanizing the Information Age, Aizu Wakamatsu City, Japan, August 25−28.

Wilks Y. (1978). *Making Preferences More Active*. Artificial Intelligence, 11/3, pp.197−223.

The collection and use of a descriptive corpus for the study of musical effect

Dave Billinge, Department of Creative Technology, University of Portsmouth, Buckingham Building, Burnaby Road, Portsmouth PO1 3AE.
Email dave.billinge@port.ac.uk

Abstract

This paper describes the structure and methods of a series of experiments carried out to study the use of figurative language in the description of musical effect. As such it focuses on practical as much as theoretical issues, these being discussed extensively elsewhere. It was first necessary to collect a set of descriptive words, refine the set by usage frequency and then analyse responses to the further use of this refined set. As a direct result of these exercises it was decided to extend the study to cover consideration of word groups. The author critically reviews the methodological processes chosen. For a review of experimental outcomes and further theoretical discussion see Billinge (2001) and Billinge and Addis (2001, 2002a, 2002b, 2003).

1. The Aim of the Experiments

Given the extent of shared informal talk and informal writing about the experience of musical performance, it was felt worthwhile to attempt a confirmation that listeners successfully communicate their feelings and to clarify which mode of linguistic communication they use. If successful the results could contribute to the creation of expert systems in artistic decision-making. Section 6 discusses briefly how theory has been revised in this respect.

2. Collecting the vocabulary

2.1 The Lexicon

The purpose of the first experiment was to identify a lexicon of descriptive words used by music lovers. To stage any experiments on the descriptions used by listeners there had to be a preliminary set of words. Lacking any previously established lexical set for such use, the choice lay between creating a list of one's own, taking it from the most easily available published source, or collecting it from the users. The creation of a list oneself was dismissed as too subject to bias. Given that any initial request for words would have to make allowance for the unwillingness of respondents in the mature age group targeted to commit themselves without guidance, it was decided that some examples had to be included. The author listed a set of words from several randomly sampled copies of *Gramophone* magazine[1]. *Gramophone* is the closest this collecting field cum hobby has to a trade paper, and revised it with the help of a colleague so as to remove at least some of the personal bias. The questionnaire included a list of 50 musical compositions. Knowing the sensitivity of the music lover to assessment of his or her knowledge of the orchestral repertoire this list was not a simple matter. It could not be all popular music because of the bias that would place on the type of music listed. The popular classical repertoire is largely late Classical and Romantic music, and promoters tend to avoid contentious or "difficult" music because concert promotion is a commercial act. The chosen list therefore had to stray a little away from this central repertoire without making any respondent feel ignorant by asking for reactions to pieces of music of which they had never heard.

The instructions included the following sentences.

If, for any one piece, you cannot think of any words, then please refer to the list on the back page for inspiration, but I am much more interested in your own words. It is quite possible that your words are in my list already; this does not matter in the slightest! Finally, it does not matter if you use the same words several times, the order in which you enter words does not matter and "Word 1, 2, and 3" are only there to guide you.

---

[1] Gramophone magazine is currently celebrating its 80th year of publication. Since 1923 it has published a monthly review of primarily classical music recordings. It is thus seen as the most important international publication of its kind.

Table 1 shows a sample of the chosen 50 musical items in the layout actually issued.

		Word 1	Word 2	Word 3
22	Ibert: Divertissement			
23	Mahler: Symphony No.2 “Resurrection”			
24	Mendelssohn: Violin Concerto			
25	Mendelssohn: A Midsummer Night’s Dream			
26	Mozart: Symphony No.40 in G minor			
27	Mozart: Eine Kleine Nachtmusik			
28	Mussorgsky/Ravel: Pictures at an Exhibition			
29	Nielsen: Sinfonia Espansiva			
30	Prokofiev: Lieutenant Kijé			

Table 1: First Questionnaire (extract)

The list of words from *Gramophone* was appended so that no one need feel unable to give a response. This was also for the purposes of keeping the respondents cooperative because many were needed for subsequent work.

impulsive	labyrinthine
individual	lacklustre
inspired	light
intricate	lightweight
inventive	lively
involved	long breathed
inward	loving
joyous	luminous
keen	lurid
kitschy	lusty

Table 2: Given vocabulary (sample)

It was hoped that this approach would result in an experimental corpus that had high user acceptability. Word frequencies were used to reduce the resulting set of 1032 words to a manageable size. Words appearing less than six times were not utilised because in common with all word usage distributions (Zipf 1949) the numbers of words repeated just a few times are huge. It is as the rate of repetition rises that the items appearing with such frequencies grow smaller. Six was chosen as the cut off point because the number of words repeated three times (33), four times (22) and five times (13) were much larger than the number repeated six times (only 8) and would thus have made membership of the “common” set skew disproportionately to less frequent words.

2.2 The Respondents and the Responses

The 12 respondents volunteered from a group of about 60 attending a music day school. This initial group was smaller than intended because of administrative problems with the organisers who, oddly, considered the author’s questionnaire an attempt to use their customers without their permission. This is mentioned here in a methodological discussion as a warning to those focussed on what they see as an innocent academic pursuit that not all those involved necessarily see it that way. Later experiments were better prepared in this respect and much higher responses achieved, thus any restrictions inadvertently applied to the initial vocabulary set was overcome subsequently.

No attempt was made to gain a balance by ages or sex because of the profile of attendees and the unfortunately restricted size of the group. Those involved here and later throughout the study were

representative of the local concert-going public in that they tended to be middle-aged or elderly rather than young, though youth was not actively excluded. The author bore in mind the possibility that age, education and sex might be significant in an exercise so closely allied to vocabulary size but possibly because the sample was too small, no differences arose in respect of these characteristics.

Finally in respect of personnel it should be noted that the author achieved insightful and instantaneous feedback from one volunteer who said that a request for three discrete words was not nearly enough to allow their feelings to be expressed. "Simplistic if not positively foolish" was the phrase actually used. In retrospect this pinpointed not so much an experimental design flaw as a weakness in theory subsequently acknowledged as this research has moved away from the discrete lexicon to embrace phrasal, figurative language.

To avoid loss of potentially valuable data several non-lexical facts were recorded. It was not known whether vocabulary would vary by sex or age so this was recorded. Though respondents were explicitly instructed to ignore word order (see Table 1 above) this positioning was recorded so that account could be taken if later analysis implied it to be important.

A certain amount of personal judgement and editing was also needed. Some words were used incorrectly. For example Bartók's *Music for Strings Percussion and Celesta* is not *atonal* but was so described. Despite being factually incorrect the word was admitted as a figurative usage. Word misuse had to be considered, for example it was decided that *emotive* probably meant *emotional.* Such errors were simply corrected. One respondent noted in the margin that by *varied* she meant *serious to romantic* and *amusing, dramatic.* This way of getting in more than the required number of words was accepted and the words added to the tally.

3. The First Group Experiment

3.1 Data Recording

The first group experiment utilized the above vocabulary set to explore the extent of user agreement on musical predication. Nineteen people participated in three groups on different dates including one with markedly younger testees. The questionnaire asked for a few personal details and the results anonymised and summarised as in Table 3.

Sex	Age	Occasional Listener	Regular Listener	Frequent Listener	Instrumental Player	Participant Number
F	49			X	Y	1
M	56		X		N	2
F	44	X			Y	3

Table 3: Participant Data (extract)

Sex was noted, as above, because it was possible there would be differences in vocabulary choice between men and women. Secondly the participant's age was noted. There was no evidence to support the prediction that older and younger people would choose from different vocabulary sets but it could not be excluded. The third question concerned a grading of listening experience from "Occasional" to "Frequent" listener. The assumption here was that the more experienced listener would be more likely to have read promulgations of this vocabulary in the journals and newspaper sections devoted to it and thus been influenced more. As it turned out the agreements detected were so small that such subtle analyses were redundant at best. Finally it was asked if the participant played an instrument. Since the aim of this research was the investigation of non-technical language it seemed sensible to assume that knowledge of the technical vocabulary would be influential.

Prior to commencing the experimental sessions it was emphasized that participants should not discuss anything with their fellows until specifically asked to do so. The purpose of the session was explained and that the discussions after the fourth test of this session would be recorded. At the end of the tests the composers and titles were revealed because everyone wanted to know "the answers". To avoid any

future bias no comment was passed by the author, beyond expressions of satisfaction, to indicate what he thought about the discussions he had heard.

This was the first opportunity to record a real linguistic corpus actually focused on the experimental task, communicating feelings about music in the focal language. As such this was expected to be valuable. It has been the authors experience that effort put into the quality of the recording medium is quickly repaid. It is easy to make a bad recording in which potentially vital data is lost through inadequate signal to noise ratios, background disturbance or even recording pitch instability. Hidden microphones are very unlikely to pick up subtle inflexions because they are too remote from the speaker. Experimental participants on these occasions were told that recording would be made and a good quality stereo microphone was hung, studio style, over the meeting table. Levels were checked beforehand on a professional standard cassette machine[2] (today we should use a digital medium) and the much-abused Dolby noise reduction was correctly applied on high quality tape. This effort paid off because every nuance of some prolonged conversations could now be used and reused without the effort of listening being at all burdensome. A considerable amount of the most valuable figurative language was used in quiet asides between participants, all captured clearly on tape. Whilst not exactly a methodological issue, the author believes investigators of natural language usage overlook the issue of fidelity in audio recording at their peril.

Similarly the music itself was well reproduced on a high quality system. The author reasoned that if emotional responses were sought then it would be better to ensure that there were as few distortions of reality as possible to clear the way for that reaction.

3.2 The Exercises

The sessions were divided into four exercises. Each of these short "tests" were designed to place increasing pressure on the participants to agree, culminating in an explicit demand for agreement. Each test description is followed by a short rationale.

Test 1 required each participant to listen to ten short items of classical music, mostly extracts, and write down a one word descriptive response without discussion. This produced a freely chosen list without the influence of others. Sufficient time was given for all to finish without pressure being applied.

Test 2 consisted of a replay of the same items but this time the participant had to chose one word from a given list. This provided the experimenter with a set of results that, because the range of words was restricted, had an increased likelihood of displaying agreement. Again no discussion was allowed.

Test 3 consisted of ten new sections of music with a given accompanying descriptive word. Participants were asked to say whether they agreed / disagreed on a five-point scale with the appropriateness of the given word. This test enforced even more restrictions in that just a single word was available and only its appropriateness had to be decided. Agreement on this was maximally likely short of explicit agreement, which was disallowed by the no-discussion rule.

Finally test 4, which was audio recorded, presented the group of five or six people with five slightly longer pieces with the instruction to agree a set of three appropriate words from the given list. The pressure here was to necessity of reaching agreement before the test proceeded to the next item. No time restrictions were imposed. The discussions lasted between five and fifteen minutes. This provided the author with tape-recorded evidence of the strategies adopted by a group of people negotiating their way to agreement.

4. Analysis of the Vocabulary

The objective of this analysis was to assign classes of use to descriptive words independently both from particular pieces of music and from other members of a test group. The experiment sought to analyse first the usage of the vocabulary to describe distinct categories of musical experience and second to assess the vocabulary in its capacity to convey a range of positive to negative evaluations. The researcher chose the former categories after discussion with professional musicians. These categories

[2] Sony DM6 Walkman Professional

were: value, for example the greatness of the piece; speed; mood, sad or happy etc.; tunefulness; and finally rhythm.

A 160-item questionnaire was constructed and issued to 58 volunteers. See Table 4 below. To make the analysis easier and the task of completion quicker the boxes only had to be ticked. Participants reported taking upwards of two hours to complete this, making the almost 100% return quite remarkable.

Word	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**
category of word	rhythm	tunefulness	mood	speed	value	don't know	very positive	positive	neutral	negative	very negative
sympathetic											
fluent											
forceful											
polished											
pastoral											
lacklustre											
light											

Table 4: The structure of the main vocabulary survey

Optical Mark Reading technology was used to create coded versions of responses. A small segment of the OMR output is reproduced below in Table 5.

sympathetic	fluent	forceful
nnnnnFnnnnn	AnnDnnnHnnn	AnCDnnnHnnn
nnnnEnnHnnn	nnnDnnnHnnn	AnnnnnGnnnn
nBCnnnnHnnn	nBCnnnnnInn	AnnDnnnHnnn
nnCnnnnHnnn	AnnDnnnHnnn	nnCnnnnHnnn
nnCDnnnHnnn	ABnnEnnHnnn	AnCnnnnnInn
nBCnEnnHnnn	AnnDnnnnInn	nnnDnnnnnnK

Table 5: Sample OMR output

Any boxes ticked in Table 4 by each participant are reflected in Table 5 by an upper-case letter A to K with all other cells labelled with an "n" for Not filled. These spreadsheets were exported as comma delimited files into relational database management software so that SQL (*Sequel*) interrogations could be used allowing counting, alphabetic sorting, string chopping and fuzzy searching. Thus exact matches and, most crucially, similar patterns could be found using the SQL "*like*" function. Statistics can also be derived showing the extent and size of agreement. 308 different patterns were found amongst the total 9280 submitted. Many analyses became possible with this technique but the most interesting for this research was the ability to speedily find the words exhibiting the handful of most commonly occurring patterns.

5. The Second Group Experiment

Groups again met in a domestic environment and using paper records and cassette tape audio the activities were recorded for subsequent analysis. It has been suggested by Sloboda (1999) that the physical situation of experiments might affect the outcome. He notes that some experiments, carried

out in laboratories, may owe at least some part of their outcome to the artificiality of the surroundings. Given that the normal surrounding for listening to recorded music is the home, the author decided to apply the normalisation as far as possible by inviting the 19 participants to his home in three smaller groupings of 6, 6 and 7. This had the further benefit of allowing the use of high quality domestic, rather than lower quality institutional, playback equipment; the actual sound of the music was easier on the ears of the participants. In addition, since the study is specifically of informal discourse, the informality can be increased by the provision of tea, biscuits, wine, sandwiches etc. This is not a trivial issue. Most informal discussion of music takes place between concertgoers in bars during intervals or in bars after the concert. If anything the tea is the unreality.

For the experimenter the music to be played was listed along with a predicted word set as shown in table 6. This set was predicted after discussion with a colleague to act as a baseline for further analysis. As noted below it turned out to be no more than an experimental whim, like some sort of minority report.

Item Sequence	**Music (circa 1 or 2 minutes)**	**Vocabulary Set**
1	Prokofiev: Alexander Nevsky; The Battle on the Ice (opening) Decca CD 410 164-2 Track 1	icy, tense, crystalline, glacial, graphic
2	Milhaud: Scaramouche (3rd movt.) DG LP 2531 389 Side 2 Track 3	vivacious, lively, joyous, rollicking, spirited
3	Weill: Surabaya Jonny : Happy End (opening) DG LP 2563 585, Side 1, Track 9	grieving, poignant, sentimental, passionate, theatrical

Table 6: Second Group Experiment: musical extracts
and predicted associated vocabulary set (extract)

The participants were given only the word sets and no indication of the music chosen. All discussions were recorded on audiotape as previously noted. The data was collected subsequently from those tapes so as not to disturb the listening environment with pauses for the researcher to finish note taking.

It was predicted that with just 15 extracts and 15 lexical sets the likelihood of agreement was being maximised. Table 7 shows the way in which data was recorded for comparison. It also shows a little of the continued lack of agreement. The darker shading (in red on the original) highlights the final decisions of groups and the greyed out highlights show the word sets also discussed. This matrix consolidated the experimental record keeping in a way the author found useful for later communication with fellow researchers

Item Sequence	Music	Vocabulary Sets **Discussed** & **Agreed**														
1	Prokofiev: Alexander Nevsky; The Battle on the Ice (opening)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
The second group discussed 8 for sometime before deciding to agree on 11. All groups discussed 8 and 10 as possibilities but there was no final agreement. The predicted choice was 8.																
2	Milhaud: Scaramouche (3rd movt.)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
All three groups discussed 3 but just two of the groups finally agreed on it as the prime descriptor set. The predicted choice was 7.																

Item Sequence	Music	Vocabulary Sets **Discussed** & **Agreed**														
3	Weill: Surabaya Jonny (Happy End) (opening)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
A rare agreement on the predicted choice, 9, and by all participants. Interestingly the second group never discussed any other set whereas the others ranged widely.																

Table 7: Second Group Experiment: agreed vocabulary set attachments (extract)

This final experiment sought to find agreement between people when the language was restricted and the music was kept within prescribed categories. The extent of disagreement was extreme with only two of the 15 extracts gaining agreement in all three groups and even then only one matched the predicted choice, making the "predicted choice" an irrelevance. The range of vocabulary discussed was very wide.

The author was led to the conclusion that, in essence, listeners do not agree in their predications, at least when a discrete vocabulary is imposed. It was discussion of this point that led to the theory being revised. Work is now ongoing that focuses on figurative, and therefore mainly phrasal structures. Initial results seem to suggest this is going to be more fruitful (Billinge and Addis 2003). The original aim of an artistic decision support system now seems more distant and possibly less interesting.

6. Methodological Conclusions

Earlier studies of the language of musical effect (for example: Gundlach 1935, Hevner 1936, Gabrielssohn 1973) were not clear about the procedures used to compile an initial descriptive vocabulary or about the approach taken to analysis of the corpus. The selection, mainly, of musically interested participants and the decision to have no control group might need consideration but the nature of the results did not imply that this approach was mistaken. The author believes that a control group would be unlikely to share the lexicon sufficiently to contribute. The focus of the research is on the means of communication. Those not sharing the language of "music talk" would fail to communicate and thus contribute no useful data. The decision to derive the initial lexical set from a mixture of personal knowledge and published sources was satisfactory because extensive user input allowed a means of refinement that gave an acceptable lexical set from the user's viewpoint. The subsequent use of reduced and multi-element lexical sets provided subjects with a more common vocabulary and hinted at the need to extend this research into tropic communication where most figurative language is phrasal rather than lexically discrete. The use of a domestic environment for group meetings seemed to encourage verbal exchange in a way not reported by other researchers. The author has now accumulated many hours of natural conversation as well as substantial paper records. This corpus remains valuable despite a fairly drastic revision of theory (Billinge and Addis 2001, 2002a, 2002b). The tapes are currently being reanalysed to provide input data for experimental models of this inferential mode of discourse (Billinge and Addis 2003).

References

Billinge, D. (2001). *An Analysis of the Communicability of Musical Predication* Unpublished PhD Thesis: University of Portsmouth.

Billinge D. & Addis, T. (2001). *Some Fundamental Limits of Artistic Decision Making* Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science: The Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

Billinge D. & Addis T. (2002a) *Modelling the Role of Metaphor in Artistic Description*: ECAI 15th European Conference on Artificial Intelligence, Lyon, France. Workshop 17 Creative Systems: Approaches to Creativity in AI and Cognitive Science pp 55-58

Billinge D. & Addis T. (2002b) *Towards Constructing Emotional Landscapes with Music*: in George S. (Ed) The Visual Perception of Music Notation (US publication pending IGP 2003)

Billinge D. & Addis T. (2003) *The Functioning of Tropic Communication: A Mechanism for Consistent Figurative Descriptions of Artistic Effect*. AISB'03 Symposium on AI and Creativity in Arts and Science

Gabrielsson, A., (1973) *Adjective ratings and dimension analyses of auditory rhythm patterns*. Scandinavian Journal of Psychology 14, pp.244-260

Gundlach, R.H. (1935) *Factors determining the characterization of musical phrases*. American Journal of Psychology 47, pp.624-43

Hevner, K., (1936) *Experimental studies of the elements of expression in music*. American Journal of Psychology 48, 246-268

Sloboda, John A (1999) *Everyday uses of music listening: a preliminary study*; in Yi, Suk Won (1999) Music, Mind and Science. Seoul National University Press. Korea

Zipf, G (1949) Human Behaviour and the Principle of Least Effort. Addison Wesley, London

The Corpora of Mandarin Chineseand German Animal Fixed Expressions: A Cognitive Semantic Application*

Shelley Ching-yu Hsieh,
Department of Applied English
Southern Taiwan University of Technology
Taiwan, R.O.C.
shelley@mail.stut.edu.tw

1. Introduction

This paper presents the results of a cross-lingual study of a Mandarin Chinese (MCh) corpus and a German (Ge) corpus of fixed animal expressions (AEs). The AEs in the corpora include: metaphors, similes, proverbs, sayings, frozen collocations, grammatically ill-formed collocations and routine formulae, all of which are fixed expressions (Alexander 1978, Carter 1987, Moon 1998), not ad-hoc terms or freely generated phrases, and contain at least one animal name that has metaphorical meaning. The Chinese corpus contains 2980 and the German corpus 2630 written and spoken AEs. The data are categorized by the animal names in alphabetical order in EXCEL. Different kinds of data relating to individual AEs were recorded in up to 12 separate fields, including phonetic transcription of the MCh, word-to-word translation, semantic feature of the metaphorical vehicle (animal name), frequency, metaphorical tenor (meaning) of the AE, the underlying conceit (the association between vehicles and tenors), etc.

The purpose of this research is first to examine the underlying conceit then the metaphorical tenors of these expressions in both languages. I discuss the proportions of different types of underlying conceits and the salient metaphorical tenors they convey, and finally the focus is set on the positive and negative tenors which bring out the last result that AEs are our vocabulary of values.

Over the years there has been continuing interest in the research of idioms, metaphors and in recent years also the cognitive endeavors. In comparison, studies on animal expressions are relatively few. Brinkmann (1878) investigates AEs in English, German, Italian, Spanish, French and Portuguese with the focus on domestic animal names, then Riegler (1907) completes Brinkmann's collection with wild animal names. They both study the origins of the AEs and Riegler also reports the etymology of the animal names. Craddick and Miller (1970) examine the animal names used to represent outer or inner circle for men and women and have their identification of the concept of self in terms of animal metaphors. Fraser (1981) examines insulting terms using animal names in eleven languages. The aim is to inspect if the informants have equivalent usage in their native languages as the English stupid-donkey, coward-chicken, sneaky-snake, mean-dog, nasty-rat and dirty-pig. One of the results

* The research reported here as a part of the results of the project on *A Semantic and Pragmatic Study on Metaphors of Created Animals in Mandarin, Taiwanese, English and German* would have been impossible without the support of National Science Council (NSC 91-2411-H-218-003) in Taiwan.

shows that stupid-donkey and dirty-pig are more widespread while nasty-rat is not.

Whaley and Antonelly (1983) reveal the assumptions about male-female relationships by animal metaphors; in particular the *women are animals*. According to Low (1988) and Newmark (1988), animal metaphors are largely used to describe inferior or undesirable human habits and attributes. O'Donnell (1990) lays his focus on the descriptions of common and productive figurative meanings assigned to animal names and animal metaphors in different languages. Sutton (1995) studies language discrimination towards females and makes a strong argument on *women are animals* metaphor. Hsieh (2002) further discusses animal expressions in light of the approach of semantic molecules (Goddard 1998). She suggests the interconnection and interaction between semantic molecules and these animal names serve as semantic contributors in distinct semantic domains. Fontecha and Catalan (2003) concentrate on the word pairs *fox*/*vixen* and *bull*/*cow* and their Spanish counterparts *zorro*/*zorra* and *toro*/*vaca* with the data from dictionaries to investigate the semantic derogation of the related animal metaphors and concepts. They found that, with mapping from source to target domain, the main metaphorical meanings of the female terms connote worse qualities than those connoted by the metaphors of the male terms.

2. The underlying conceit

As Lakoff and Turner (1989:65) already noticed "We conventionally understand these concepts not by virtue of metaphoric mappings between them and different conceptual domains but rather by virtue of their grounding in what we take to be our forms of life, our habitual and routine bodily and social experiences." Most of the AEs reflect human observation of the vehicles. Both Chinese people and Germans may observe and perceive animals from the same viewpoint and interpret what they see identically. i.e. they share the same underlying conceit. For example, the ease of fish in water is expressed in MCh as *ru^{2}yu^{2}de^{2}shui^{3}* 如魚得水 (as-fish-get-water – feel just like fish in water; be in one's element) and in Ge as *wie ein Fisch im Wasser* (like a fish in water – feeling well).

Both peoples may share the same viewpoint but develop different underlying conceits and therefore generate different AEs, e.g. the cat is gluttonous in the eyes of both Germans and Chinese, thus *nei^{3}zhi^{1}mao^{1}er^{2}bu^{4}tou^{1}xing^{1}* 哪隻貓兒不偷腥 (which-cat-not-steal-raw-fish – which cat wouldn't steal the fish smell; which man wouldn't like the wife of another) developed in MCh. The German version is *Naschkatze* (sweet-toothed cat), which means a person who likes nibbling at sweets. Both emphasize human behaviours, but the MCh is in the domain of 'emotion' while the Ge belongs to the domain of 'basic need'.

AEs are developed either from the animals' appearances, habits, and relation to people (Wierzbicka 1985:167) observed from different cultural backgrounds. In addition, many AEs are arbitrary inventions and have nothing to do with the animals themselves (Hsieh 2001:149-), as exemplified in Tables 1 and 2: 15% in MCh and 9% in Ge. Without doubt, most of the underlying conceits of AEs in both languages are associated with the metaphorical vehicles' attributes, e.g. their appearances, habits or behaviours. Ahrens and Say (1999:6) propose that the appearance of an animal

is usually mapped to the target domain of human appearance in MCh AEs, whereas animal behaviours are mapped to human behaviours. The result of the present corpora further indicate that Chinese tend to generate more AEs from animal appearances and apply them to the basic-need domain (see Table 1), e.g. that a snail carries a shell is observed by Chinese people, thus, *wu²ke²gua¹niu²* 無殼蝸牛 (no-shell-snail – people who are not capable of purchasing houses) and *gua¹niu²zu²* 蝸牛族 (snail-tribe – people who do not possess real estate) are produced, to apply to the basic housing need. On the other hand, the Germans tend to generate more AEs from animal behaviours or habits and apply them to an emotional domain (see Table 2). That a snail carries its shell is also observed by the Germans, but the behaviour that it withdraws into its shell when encountering danger is the conceit of the AEs: *sich in sein Schneckenhaus zurückziehen* (self-in-one's-snail shell-withdraw) and *jemanden zur Schnecke machen* (someone-to-snail-make) They are composed to denote "to go into one's shell" and "to come down on someone like a ton of bricks", respectively. Tables 1 and 2 count the percentages of different types of underlying conceits and the share of metaphorical tenors in MCh and Ge.

Table 1. The underlying conceits and metaphorical tenors in Mandarin Chinese corpus

Underlying Conceit	**Percentage**[1]	**Metaphorical tenor**	**Percentage**
Appearance	27%	basic need domain	**25.8%**
		emotion	5.1%
		amusement	5.4%
		society	**14.2%**
		work, sport, etc.	49.5%
Behavior	25%	basic need domain	**29.2%**
		emotion	11.1%
		amusement	5.5%
		society	**11.1%**
		work, sport, etc.	43.2%
Habit	18%	basic need domain	**22.2%**
		emotion	9.5%
		amusement	5.1%
		society	**16.0%**
		work, sport, etc.	47.2%
Human-Animal Relation	21%		
Arbitrary	15%		
Unknown	8%		

The unknown derivation as shown in the tables can be traced from historical events and be arbitrary inventions. The popular Ge AE *Mein Name ist Hase* (my name is hare – I know nothing; search me) is an example: At the end of the semester 1854/55 Victor von Hase helped a student illegally cross the German boarder by providing him with his own identification passport. As he was interrogated by the police he replied immediately, "My name is Hase (hare), I deny all questions, I

[1] An AE can be categorized into more than one type when we analyze its underlying conceits, e.g., *qian¹xi¹chong²* 千禧蟲 (millennium-bug – y2p; year 2000 computer problem) can be associated with the small size of the bug – appearance, and the harm that it brings – habit. Therefore, the total percentage of underlying conceit in Table 1 is 110%, and that of Table 2 is 105%.

know nothing at all." This statement went the rounds in Heidelberg and became a well-known saying from then on (Büchmann 1937:579). Arbitrary inventions are mostly abstracts of legends and superstitions. They can be due to rhyme form, e.g. *weder Fisch noch Fleisch* (neither fish nor meat – neither fish nor fowl; neither ass nor horse; ambiguous). Or like many modern AEs, e.g., (transliteration) *ma^{3}sha^{1}ji^{1}* 馬殺雞 (horse-kill-chicken – transliteration of English "massage") and (phonetic translation) *ma3ke4* 馬克 (horse-gram – Deutsche mark). Language contact brought out more and more such inventions.

The metaphorical vehicles fish, dog, horse, mouse, etc. generate AEs based on the vehicles' habits, as in the above exemplified *wie ein Fisch im Wasser* (like a fish in water – feeling well). Their AEs also often are based on human-animal relations (fishing, watchdog, horse riding, culture follower). This is a marked feature of more productive vehicles. Less productive vehicles tend to render specific underlying conceits and generate particular metaphorical tenors, such as *mao* 貓 (cat) for 'gluttonous' and *Kater* (tomcat) for 'hangover'. The domains of metaphorical tenors will be discussed in the following section.

Table 2. The underlying conceits and metaphorical tenors in German corpus

Underlying Conceit	**Percentage**	**Metaphorical tenor**	**Percentage**
Appearance	21%	basic need domain	10.6%
		emotion	8.4%
		amusement	8.4%
		society	9.0%
		work, sport, etc.	63.6%
Behavior	27%	basic need domain	10.9%
		emotion	**13.5%**
		amusement	4.3%
		society	3.6%
		work, sport, etc.	67.7%
Habit	21%	basic need domain	9.8%
		emotion	**14.5%**
		amusement	4.4%
		society	5.7%
		work, sport, etc.	65.1%
Human-Animal Relation	20%		
Arbitrary	9%		
Unknown	12%		

3. The metaphorical tenor

First the salient metaphorical tenors will be distinguished, then the evaluation of some tenors in order to represent the different values.

3.1 Salient metaphorical tenors

When I examine the underlying conceits that belong to animal attributes – appearance, behaviour and habit, both Tables 1 and 2 indicate that MCh and Ge favor the metaphorical tenor of the BASIC NEED domain in which they utter the various meanings about eating, drinking, housing, etc. Lakoff and Turner (1989:168) said in their "Great Chain Metaphor" that "Therefore, instinct is a generic-level parameter of animals. Similarly, the mental, the moral, and the aesthetic are generic-level parameters of

human beings."

In addition, Chinese people tend to create more tenors related to the SOCIETY domain while Germans ring the bell for the EMOTION domain.[2] There are a good number of group-oriented secular benedictions in Mandarin Chinese and many endearments (one-on-one dictions) in German, but not vice versa. For example, *wo*^{4}*hu*^{3}*cang*^{2}*long*^{2} 臥虎藏龍 (crouch-tiger-hide-dragon – a remarkable talent who has not been discovered), *Schmusekatze* (flattering cat – a term of endearment to a woman). This gives a hint to the different modes of thinking between Chinese and German, i.e. the Chinese tend to think group-centrically while the Germans think individualistically or egocentrically (Hsieh 2002). On the other hand, German endearments fall into the EMOTION domain, while the MCh secular benedictions express the SOCIETY domain. A SOCIETY domain like schooling can be exemplified by the AEs *ren*^{2}*sheng*^{1}*bu*^{4}*du*^{2}*shu*^{1} *huo*^{2}*zhebu*^{4}*ru*^{2}*zhu*^{1} 人生不讀書 活著不如豬 (people-life-not-read-book-live-not-as-pig – people living in the world would be ignorant if they did not study), *fang*^{4}*niu*^{2}*ban*^{1} 放牛班 (release-cow-class – let alone classes where the students' school performances are inferior), *ya*^{1}*dan*^{4} 鴨蛋 (duck-egg – the school grade "unsatisfactory": zero), *shang*^{4}*ke*^{4}*xiang*^{4}*tiao*^{2}*chong*^{2} *xia*^{4}*ke*^{4}*xiang*^{4}*tiao*^{2}*long*^{2} 上課像條蟲 下課像條龍 (up-class-like-a-worm down-class-like-a-dragon – students acting dully in class and dynamically out of class) and *diao*^{4}*yu*^{2} 釣魚 (fishing – sleepy; to fall asleep for tiredness in class). Chinese also emphasize diligence as a human virtue, such as with *wen*^{2}*ji*^{1}*qi*^{3}*wu*^{3} 聞雞起舞 (hear-chicken-up-dance – to rise up upon rooster; diligent and full of enthusiasm) and *li*^{1}*ba*^{1}*za*^{1}*de*^{2}*jin*^{3} *huang*^{2}*gou*^{3}*zuan*^{1}*bu*^{2}*jin*^{4} 籬笆紮得緊 黃狗鑽不進 (fence-basketry-tie-get-tight, yellow-dog-drill-not-inside – man should work hard to prevent a contingent disaster). However, diligence is not emphasized in a German-speaking society.

Some examples from the German EMOTION domain, other than the above-mentioned German endearments, are: *einen Affen an jmdm. gefressen haben* (to have eaten a monkey on someone – to be crazy about someone), *Du benimmst dich wie ein Backfisch* (you behave like a fried fish – you behave like an young girl falling in love), *jmd. umklammern wie ein Tintenfisch* (someone embrace like a squid), *einen Vogel haben* (a-bird-have – to have a screw loose) and *die Sau rauslassen* (the-sow-let-out – to let the pig out; having fun; to paint the town red).

When categorized, the metaphorical vehicles horse, dog, cow, etc. tend to be responsible for 'work', pig, snail, etc. more for the BASIC NEED domain 'housing', and the names of wild animals more for SOCIETY. There are vehicles that serve only as positive metaphorical tenors, such as *long* 龍 (dragon) in MCh. Many vehicles produce only negative metaphorical tenors, such as *gou* 狗 (dog) and *Hund* (dog). Some vehicles serve for specific metaphoricality, such as German *Grille* (cicada) stands for 'strange mood' and 'strange ideas'. Moon (1998:163) says that "idioms represent concepts embedded in the culture and associated with particular lexicalizations. They are characterized by an underlying conceit … and an overlying preferred lexical realization", and usually with connoted evaluation. The

[2] The tenors of the terms of endearments are categorized into the BASIC NEED domain "love" that can also be sorted to EMOTION.

present corpora demonstrate that about 80% of AEs are used to scorn or warn people. Thus, we can say AEs are a vocabulary of peoples' values. They convey values from different cultures and societies. The following sections exemplify this argument.

3.2 Positive and negative tenors

Both Germans and Chinese pay attention to their shape and watch their weight. Praises in forms of AEs are: *shui^{3}she^{2}yao^{1}* 水蛇腰 (water-snake-waist – a slender waist), *Wespentaille* (wasp waist – slender waist), *shen^{1}qing^{1}ru^{2}yan^{4}* 身輕如燕 (body-light-like-swallow – light as a swallow) and *schlank wie ein Reh* (slender-like-a-deer – slender). People outside the norm are despised with AEs like *fei^{2}zhu^{1}* 肥豬 (fat-pig – a fat person; fatty), *shou^{4}pi^{2}hou^{2}* 瘦皮猴 (thin-skin-monkey – bag of bones), *Schwer wie ein Elefant* (heavy like an elephant – very heavy), *Schultern wie ein Huhn* (shoulders like a chicken – having slim shoulders), *pudeldick* (poodle fat – very fat) etc.

Table 3 gives the evaluation of body-part AEs. Although many of them are neutral descriptions, such as *hu^{3}kou^{3}* 虎口 (tiger-mouth – part of the hand between the thumb and the index finger) and *tu^{4}chun^{2}* 兔唇 (hare-lip – harelip; cleft lip), some of them are compliments, most of them carry negative connotations.

Table 3. The evaluation of body-part AEs

Evaluation	Mandarin Chinese	German
positive	13%	18%
negative	66%	64%
neutral	21%	18%

Also to pinpoint and reprove a woman are examples, *aufgedonnert wie ein Pfau* (in full feather like a peacock – dressed or done up to the nines) and *hu^{2}li^{2}jing^{1}* 狐狸精 (fox-spirit – woman of easy virtue [supposed to be a fox in disguise]; an enchantress),. To a man, e.g., *alter Gockel* (old cock – a conceited old man) and *wu^{3}ye^{4}niu^{2}lang^{2}* 午夜牛郎 (mid-night-cowboy – male prostitute). To events, e.g., *hua^{4}she^{2}tian^{1}zu^{2}* 畫蛇添足 (draw-snake-add-foot – draw a snake and add feet to it; ruin the effect by adding something superfluous) and *Schweinearbeit* (pig work – chore). To places, *bei euch ist ein furchtbarer Hecht* (there is a terrible pike in there – there is a stale air in there) and *gou^{3}wo^{1}* 狗窩 (dog-den – doghouse; small and in disorder room). And to a society, e.g., *die großen Fische fressen die kleinen* (the-big-fish-eat-the-small – the great fish eat up the small; the strong overwhelm the weak) and *shu^{4}dao^{3}hu^{2}sun^{1}san^{4}* 樹倒猢猻散 (tree-fall-monkey-scatter – when the tree falls, the monkeys scatter; when an influential person falls from power, his hangers-on disperse).

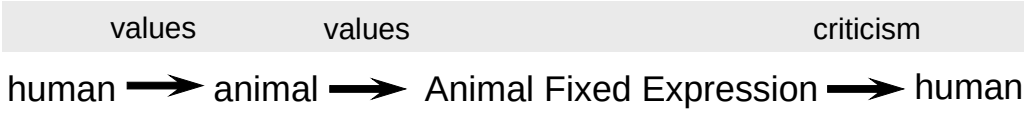
A great amount of AEs are taboo, e.g., *Sauigel* (sow-hedgehog – a person telling indecent jokes; dirty person), *dummes Kamel* (stupid camel – stupid!), *er geht nicht mit kleinen Hunden pinkeln* (he does not go pissing with small dogs – he is not interested in insignificant people),

lang²dao⁴chu⁴chi¹rou⁴ *gou³dao⁴chu⁴chi¹shi³* 狼到處吃肉 狗到處吃屎 (wolf-everywhere-place-eat-meat, dog-everywhere-place-eat-dung – people in different classes have different lives), *shen¹ru²lan⁴chan²* *cui³ru²tie³qian²* 身如爛蠶 嘴如鐵箝 (body-like-rotten-silkworm, mouth-like-iron-tongs – to blame someone who does not admit his mistake), etc. Trudgill (1974:29-31 in Risch 1987:353) explains that "Such words are not only considered inappropriate for a certain context, but are forbidden in most communicative contexts." However, there is an underlying cognition that we adopt animal names as metaphorical vehicles and create a great quantity of AEs as part of our vocabulary.

4. Vocabulary of values

The corpora document the compliments and taboos that express the differences and the similarities between human beings and animals. AEs are used not always for bad purposes but rather due to some ignorance with respect to the nature of the animal (Schenda 1998:13).[3] In other words, the metaphorical vehicles that people adopted to produce AEs and people's knowledge of animals are often based on different cognitive levels. For example, we know monkeys are clever, but we have AE *Affentheater* (monkey-theater – complete farce) and *sich zum Affen machen* (make a monkey of oneself – to make a fool of oneself). Zoological research (Grzimek 1988:20 and elsewhere) reports that pigs are smart, but *ben⁴zhu¹* 笨豬 (dumb pig; idiot) is a popular AE. People use AEs as swearwords and as emphatic comparisons as Michel (1991:ii) states: … the silly donkey and the sharp-eyed falcon. We human beings imagine ourselves as above other animals because animals are merely controlled by their instincts. Nevertheless, we also envy other animals because of their excellent senses and abilities.[4]

Fig. 1 AE schema



AEs express positive and negative sanctions in the societies. Praise and reprimand help the process of adaptation to the standards and rules of the society. When one is called a *falscher Hund* (a false dog – a false man; a liar), he should know that his behaviour is considered to be "false, underhanded, insidious" and should change his attitude accordingly. When being called a

[3] The original text is: "nicht immer aus böser Absicht, eher aus Unwissenheit."

[4] The original text is: "der dumme Esel und der scharfblickende Falke. Wir Menschen fühlen uns teils erhaben über die in ihren Erbkoordinationen befangenen Tiere; teils aber beneiden wir sie auch um ihre vorzüglichen Sinne und Anpassungen in lebensbedrohenden Umwelten."

gen[1]*pi*[4]*chong*[2] 跟屁蟲 (follow-butt-worm – bluebottle) one knows that it is improper to cling to someone like a leech. Huang and Tian (1990:83) explicate vocabulary with negative denotations as: "Modern linguistic taboo is chiefly due to regard for social etiquette, propriety in behavior … Inhibition, rather than prohibition, is the key to understanding the very intricate nature of linguistic taboos in our time."

To conclude, AEs are a vocabulary of peoples' values used to express our values and to criticize human behaviours. Fig. 1 shows the schema of the application of animal fixed expressions as human criticism or evaluation. People map their system of values subconsciously on animals and imagine how animals should be, then generate AEs accordingly, with the systematic underlying conceit and the metaphorical tenors surfaced, to criticize and to rule human beings themselves.

References

Ahrens K, Say A L T 1999 "Mapping Image-Schemas and Translating Metaphors." In: *The Proceedings of Pacific Asia Conference on Language, Information and Computation.* Taipei.

Alexander R J 1978 Fixed expressions in English: a linguistic, psycholinguistic, sociolinguistic and didactic study (part 1), *Anglistik und Englischunterricht* 6:171-88.

Brinkmann F 1878 *Die Metaphern*. Verlag von Adolph Marcus, Bonn.

Büchmann G 1937 *Geflügelte worte – der zitatenschatz des deutschen volkes*. Berlin, Haude & Spenerschen.

Carter R 1987 *Vocabulary*. London, Allen and Unwin.

Craddick, Miller 1970 Investigation of the symbolic self using the concentric circles method and animal metaphor. *Perceptual and Motor Skills*, 31:147-150.

Davies E E, Bentahila A 1989 Familiar and less familiar metaphors. *Language & Communication*, 9:49-68.

Drosdowski G, Scholze-Stubenrecht W 1992 *Duden (Band 11) Redewendungen und sprichwörtliche Redensarten.* Mannheim, Leipzig, Wien, Zürich, Meyers LexikonVerlag.

Fontecha A F, Catalan R M J 2003 Semantic derogation in animal metaphor: a contrastive-cognitive analysis of two male-female examples in English and Spanish. *Journal of Pragmatics*, 35:771-797.

Fraser B 1981 Insulting Problems in a Second Language. *TESOL Quarterly*, 15:435-441.

Grzimek B 1988 *Grzimeks Enzyklopädie (Band 5) – Säugetiere*. München, Kindler Verlag.

Hsieh, S C Y (謝菁玉) 2001 *Teirmetaphern im modernen Chinesischen und Deutschen: eine vergleichende semantische und soziolinguistische studie*. unpublished PhD thesis, Tübingen University.

---. 2002 Cat Expressions in Mandarin Chinese and German – Animal Expressions and Cultural Perspectives. In *the 10th International Conference on Cognitive Processing of Chinese and Other Related Asian Languages*. Taipei, National Taiwan University Dec. 9-11.

---. 2003a Vogel metaphern und Taiwan heute: sprache und was dahinter steht. *Taiwanese culture and literature news*. Germany, Ruhr University Bochum. (paper accepted)

---. 2003b The Corpora of Mandarin Chinese and German Animal Expressions: An Application of Cognitive Metaphors and Language Change. In *Proceedings of the Corpus Linguistics 2003 conference*, Lancaster University, England, pp 332-341.

Huang H, Tian G 1990 A sociolinguistic view of linguistic taboo in Chinese. *International Journal of the Sociology of Language* 81: 63-85.

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

Lakoff T M 1989 *More than cool reason: a field guide to poetic metaphor*. Chicago, University of Chicago Press.

Low G D 1988 On teaching metaphor. *Applied Linguistics*, 9(2): 125-47.

Michel P 1991 *Tiersymbolik.* Bern, Peter Lang.

Moon R 1998 *Fixed expressions and idioms in English*. Oxford, Clarendon Press.

Newmark P 1988 *Approaches to Translation*. Prentice Hall, Hemel Hempstead.

O'Donnell Paul 1990 Entre chien et loup: a study of French Animal Metaphors. *The French Review*, 63: 514-523.

Riegler R 1907 *Das Tier im Spiegel der Sprache*. C.A.Kochs Verlagsbuchhandlung, Dresden, Leipzig.

Risch B 1987 Women's derogatory terms for men: That's right, 'dirty' words. *Language in Society* 16: 353-358.

Röhrich L 1991 *Lexikon der sprichwörtlichen redensarten*. Freiburg, Herder.

Schenda R 1998 *Who's Who der Tiere – Märchen, Mythen und Geschichten.* München, Deutscher Taschenbuch Verlag.

Sutton L A 1995 Bitches and Skankly Hobags. The Place of Women in Contemporary Slang. In: Hall K and Burholtz M (eds.), *Gender Articulated*. Routledge, London.

Whaley C, Antonelly G 1983 The birds and the beasts – woman as animal. *Maledicta*, 7, 219-229.

Wierzbicka A 1985 *Lexicography and Conceptual Analysis*. Ann Arbor, Karoma.

A corpus-based study of metaphor in information technology

Sattar Izwaini
Centre for Computational Linguistics, UMIST, PO Box 88,
Manchester M60 1QD, UK
Sattar.Izwaini@student.umist.ac.uk

Key words: Metaphor, Information Technology, Corpus Linguistics, English.

Abstract

To cater for reference purposes and term creation, the language of information technology (LIT) has made use of items in our surroundings and borrowed them figuratively into its own domain. The present paper is based on a specialised corpus of English IT texts of more than 7 million words, built mainly from online help files of computer systems and software. First, metaphors are outlined for two fundamental elements in IT: *the computer* and *the Internet* by proposing conceptual categories of metaphoric words used in IT. Then, words that are used metaphorically in this field are accounted for by using statistical methods. Metaphoric words are found to be persistent in LIT. Most of the words in those categories are found to be key words in the corpus.[1]

1. Metaphor in information technology

Metaphor is generally known as being used in reflecting and developing scientific ideas (see Gross 1990, Rothbart 1984, Hesse, 1980). According to Richards "Literal language is rare outside the central parts of sciences" (1936). Dirven (1985) demonstrates the role of metaphor in extending the lexicon. The linguistic potential of metaphor has rendered it a very useful tool in providing description and clarification in various scientific domains. In scientific and technical vocabulary, lexical items of general language are figuratively used to form a special language vocabulary. Metaphor plays a significant role in scientific discourse and terminology and in transmitting scientific concepts especially in new fields. It is widely used in Information Technology. The type of metaphor and the tasks assigned to it in science and technology are fundamentally different from its role in literature. The figurative aspect of metaphor is utilized to forward a model to understand scientific facts, theories and concepts. At this point, metaphor in science and technology moves rather into terminology and specialised language.

Metaphor in the world of computers has attracted the attention of researchers in the fields of technical writing and human-computer interaction. The first group discussed how metaphor is used to present this field and what the criteria are to choose metaphor (e.g. Chisholm 1986; Johnson 1991; Beck 1991; Mulder 1996). The second group discussed the use and significance of metaphorical representations in the graphical user interface (e.g. Constantine 2001; Coyne 1995; Microsoft 1993; Apple 1987). Other studies investigated the use of metaphor in different fields of information technology from a cognitive/linguistic perspective (e.g. Grevy 1999; Meyer et al. 1997; Öberg 1989).

Chisholm (1986: 198) calls computer terminology used metaphorically *metaphoric terminology*. He maintains that metaphoric terminology is a special kind of metaphor and a sub-category of *catachersis*, a term used by Max Black and Colin Turbayne after Stanford to give a name to something that lacks a designation by borrowing it from another (ibid: 1986).

Johnson (1991) presents metaphors used in computer science as having paradigms: *Agent Paradigm* (the doer metaphor), *Engine Paradigm*, *Traffic Paradigm*, *Structure Paradigm* (e.g. *architecture*), and *Illusion Paradigm* (e.g. *virtual*). These are categorised on semantic sets of words used metaphorically in this field.

Grevy (1999) studied metaphors in the computer domain in Danish. He found that one sixth of those he collected (3000 metaphors) are *highway* metaphors. He also introduced the term *integrated metaphor* (*integrerede metafor*) to describe the way those metaphors work: 'they are integrated with other metaphors in the same semantic field' (ibid, 173, 199). His main categories are *Guest and Visit*, *PC Driving* which includes *kør* (to run/drive), and *Highway and Travel*.

Meyer et al. (1997) studied metaphors of the Internet from a conceptual and structural point of view. They looked at English books and magazines as well as online and hard copy dictionaries and glossaries. They classify Internet metaphors into two main groups: fully metaphorical and partly metaphorical, where members of the latter have either a metaphorical modifier or a metaphorical base, e.g. *kill file* and *electronic mail* respectively (1997:14). Actually the metaphoric aspect does not lie in

[1] I would like to thank Carlo Grevy for his comments on an earlier version of this paper. However, any shortcomings that remain are my own.

one element of these terms but rather from the combination of both. The constituent elements of these metaphorical expressions are not metaphorical when used on their own. Only when combined together do they give rise to the figurative meaning.

Metaphoric designations facilitate communication among field experts as well as presenting the components of the fields of computer and the Internet to the ordinary user. IT makes use of metaphor by having a mental model for the user through linguistic representations. In addition to vocabulary innovation, e.g. *byte*, and derivation, e.g. *computer* and *server*, metaphor is the most used method of creating new vocabulary in the language of information technology (LIT). Instead of trying to create new coinage, language users tend to make use of what is already available in the language by making figurative use of it. Metaphorical designations are based on the correspondence to items that are found in the real world and have some other nature; most IT entities are of electronic or magnetic nature. This takes us to the basic definition of metaphor: to describe one entity by the qualities of another. The difference in the material and nature of items is the basis on which this metaphor is created.

Metaphor is used in LIT in single words, e.g. *mouse*, *chip, card* (depending on the shape), *file*, *hardware*, *traffic*, *surf*, *page* and *port*, and in compounds, e.g. *search engine*. LIT vocabulary draws its metaphoric character from general language and everyday experience. LIT uses metaphor and assimilates its shaping boundaries in the terminology in that the figurative aspect is no longer felt.

2. Data and methodology

This paper is based on a specialised corpus of English IT texts of more than 7 million words built mainly from online help files of computer systems and software as well as diverse IT material such as manuals, tutorials, software reviews and IT journalistic items. The corpus also includes IT-specialised web sites (see Izwaini 2003).

The methodology is to classify metaphoric vocabulary within categories as well as to account for words of very high frequency that are used metaphorically in this field to create its terminology. The starting point is the conceptual categories of metaphors in LIT. First, conceptual metaphors are outlined for two fundamental elements in IT: *the computer* and *the Internet* (see *3. Conceptual Framework* below). Words that are used metaphorically in this field to create its terminology are accounted for by using statistical methods. I used the Wordsmith tools package (Scott 1997), which produces frequency and key word lists. A key word is a word that has unusual frequency in a text in comparison to a reference corpus. The key word list is generated by comparing the word list of the corpus with the BNC as a reference corpus. The level of keyness and frequency are taken as criteria of the usage of metaphor in LIT. Different word forms or *lemmas* of metaphoric words outlined in the categories are accounted for as well, e.g. *bug*, *debug*, *debugging* etc. including compounds such as *toolbox*, *toolkit*. In calculating lemmas, acronyms and abbreviations are considered one word form of the head noun, e.g. *RAM* and *ROM* of *memory*, and *http* of *protocol*.

Statistics took into consideration the syntagmatic and semantic relations of key words. Lemmas are looked at to see whether they collocate with *computer* and *Internet*. Metaphor is sometimes manifested in having one collocate changed or in having a new collocation that gives rise to figurative usage (Izwaini 2000: 24-25).

3. Conceptual framework

Metaphor is used to express different aspects of life and everyday activities in a systematic way. Lakoff and Johnson (1980) present a conceptual account of the metaphoric system and how is that embodied in language. However, this is not based on comprehensive empirical data and might not lead to conclusive results. Furthermore, linguistic factors play a role in creating metaphor; changing collocations or even violating them produce metaphors (Izwaini 2000: 24-25), e.g. *I've invested a lot of time in her*, where *invest* is a typical collocate of *money* not *time*.

Here, the conceptual framework is based on our classification of words used in IT, which results in categories or themes of metaphors. Taking two main components of information technology, i.e. *the computer* and *the internet*, we can see them metaphorically by grouping LIT vocabulary in categories of a cognitive character. Words that are used metaphorically and now are part and parcel of LIT are organised in semantic sets that result in principal LIT metaphors. The two main categories of *computer* and *internet* are as follows:

The Computer

- THE COMPUTER IS A LIVING BEING: *client*, *conflict*, *dialogue* (conversation between the computer and the user), *generation*, *language*, *memory*, *protocol*, *syntax*, *widow*/*orphan*, and *virus* and *bug* (it can get ill);
- THE COMPUTER IS A WORKSHOP: *download*, *equipment*, *hardware*, *install*, *load*, *template*, and *tools*;
- THE COMPUTER IS AN OFFICE: *attachment*, *desktop*, *directory*, *document*, *file*, *folder*, *mail*, *trash can*, and *wastebasket*;

- THE COMPUTER IS A BUILDING/PLACE: *architecture*, *library*, *sign in*/*log in*, *sign out*/*log out*, *platform*, *port*, *window*, and *workstation*;
- THE COMPUTER IS A SOLDIER: *combat*, *command*, and *instructions*.

The Internet

- THE INTERNET IS IN A STATE OF WAR: *password*, *security*, *war*, and *warfare*;
- THE INTERNET IS A ROAD: *bus*, *highway*, *map*, *path*, and *traffic*;
- THE INTERNET IS A BUILDING/PLACE: *access*, *address*, *firewall*, *gateway*, *sign in*/*log in*, *sign out*/*log out*, *site*, *visit*, and *wallpaper*;
- THE INTERNET IS A BOOK: *bookmark*, *browse, browser*, and *page*;
- THE INTERNET IS A SEA: *navigate*, *pirates*, and *surf*;
- THE INTERNET IS A MARKETPLACE: *ecommerce*, *emarketing*, and *eshopping*.

Although the categorisation is different, some of these metaphors correspond to categories suggested in other studies, e.g. THE INTERNET IS A ROAD corresponds to Grevy's *Highway and Travel* (Grevy 1999), and to Johnson's *Traffic Paradigm (*Johnson 1991).

4. Statistics

Two word lists are produced to reflect the make-up of LIT vocabulary: a key word list and frequency list. The first is more significant in that it includes the words that have an unusual frequency in comparison with general language and thus have an important status. The key word list includes 500 key words. We looked first at the constituents of the metaphor categories, e.g. *language* and *war*, to see what level of keyness they have. The constituents here are looked at as words first and then as lemmas to see what percentage they have. The next step is to look at other constituents that are not present in the key word list. The frequency of those words and their lemmas is calculated to see what percentage they have and to be added in the end to the percentage of the other constituents of the same category.

There are some factors that can affect the results. First, keyness of some IT words, e.g. *virus*, is negatively affected by their non-IT meanings found in the reference corpus. Second, some words are also present in a non-metaphorical sense in the corpus, e.g. *language* and thus they can make the frequency higher. However, their frequency is rather marginal. Third, regional variants cause the word to have a different format, and according to the software calculation the word can lose or get the status of being a key word, e.g. *dialogue* and *dialog* (see 4.1.1 below).

4.1 The Computer

Computer is a key word ranking 129 in the list. Many of the constituents of the suggested categories are key words. *Computer* collocates within a short span and with different degrees of collocation with the lemmas of most of the constituents of all the categories suggested. This will be presented after the statistics of every constituent being presented. On the other hand, other words that are used for the computer such as *PC* and *machine* were also looked at.

4.1.1 THE COMPUTER IS A LIVING BEING

Taking the first theme, we find that the key words are as follows with their order of keyness in brackets: *dialog* (26), *syntax* (64), *client* (365), and *protocol* (500). Another key word, *debug*, is present under *bug* which is suggested to be a constituent. *Debug* is 345 in order of keyness. *Dialog* is a key word because the reference corpus is of British English. It occurs in BNC 66 times only. The percentage of these key words to the total frequency of key words in computer metaphor categories is 15.38%. By including lemmas of non-key words such as *language*, *conflict*, *memory*, *widow*, *orphan*, *virus*, *hibernation*, *freeze*, *life*, *assistant*, and *proxy* the percentage of the total frequency of constituents of this category is 0.51% of the whole corpus.

Words such as *sleep, awake, freeze*, *client*, *protocol, virus*, *communicate*, and *proxy* collocate with *computer* and thus support this theme. For example: *sleep* occurs 82 times, 53% of them have *computer* in the L2 slot, e.g.

… put your computer to sleep and wake it up…

Sleeps occurs three times, in two of them its subject is *computer and PowerBook* (a brand name of a portable computer). *Re-awaken* occurs once with *computer* as its object. *Computer* makes 64% of the collocates of *wake* as its direct object. *Freezes* occurs 22 times, in 50% of them it has *computer* as its subject.

Proxy is an adjective collocate of *server* which is a computer. *Server* is also one of the R1 top collocates of *client*. On the other hand, *protocol* does not collocate with *computer*, but rather with *Internet*. *www* which is an acronym with *web* as the head noun is the first top collocate of *http*

(hypertext transfer protocol) which is the key word no. 156. This is due to the fact that this is the structure of internet addresses via the World Wide Web. One top R1 collocate of *http* is *server* making the metaphor to have a double function. One acronym, *httpd* (hypertext transfer protocol daemon) was found to incorporate *protocol* and *daemon*. The latter is a server.

Virus was found to collocate with *computer* in R1 but with a low frequency. *Communicate* has also low figure collocations with *computer,* but see below:

A device that enables your computer to communicate with another computer…
…speed indicates the speed at which the computer communicates with the modem.

Machine was found to have the following adjective collocates: *host*, *partner* and *single*.

4.1.2 THE COMPUTER IS A WORKSHOP

Not only one word form of the elements of this category are key words, but also other word forms as well: *install* (113), *installed* (194), *installation* (218), and *installing*, (354). Other key words are *toolbar* (46), *wizard* (155), *device* (220), *download* (230)*, template* (269), *toolbars* (296), *utility* (425), and *task* (457). Their percentage to the total frequency of key words in the computer metaphor categories is 23.85%. By including lemmas of non-key words such as *load*, *hardware* and *equipment*, the percentage of the constituents of this category is 0.83% of the whole corpus.

Installed, *devices*, *hardware* and *downloaded* are found to collocate with *computer*, which supports this theme.

4.1.3 THE COMPUTER IS AN OFFICE

The key word *file* occupies the 2nd position in the list. Other key words include *document* (28), *folder* (34), *mail* (68), and *directory* (173). Their percentage of the total frequency of key words in the computer metaphor categories is 43.51%. By including lemmas of non-key words such as *desktop*, *attachment*, *archive*, *wastebasket*, *trash can*, *recycle bin* and *equipment*, the percentage of the constituents of this category is 1.2% of the whole corpus.

Collocates such as *file*, *documents*, *documentation*, and *desktop* support this theme. It is worth mentioning that the desktop metaphor is the most known metaphor which is often referred to because of the iconic metaphor used in the software design.

4.1.4 THE COMPUTER IS A BUILDING/PLACE

Key words include *window* (104) and *login* (148). Their percentage to the total frequency of key words in computer metaphor categories is 7.28%. By including lemmas of non-key words such as *sign in/out*, *firewall*, *workstation*, *platform*, *architecture*, *port*, and *gateway*, the percentage of the constituents of this category is 0.22% of the whole corpus.

Collocations are found to include *window* only. However, it was found that *server*, which is a kind of a computer named after its function, collocates with *log*, *port*, *platform* and *storage*. Both *server* and *pc* collocate with *architecture*. *Computer* collocates with *platform* and its plural form. *Machine* was found to collocate with *architecture*, *local*, *remote*, and *firewall*. *PC* has the collocates *remote*, *client*, *host* in L1 slot. In R1 slot it has *location*.

4.1.5 THE COMPUTER IS A SOLDIER

We have two key words which are both of the same lemma *command* (39) and *commands* (146). Their percentage of the total frequency of key words in computer metaphor categories is 9.97%. By including the lemmas of the non-key word *instruction*, the percentage of the constituents of this category is 0.21% of the whole corpus. As a soldier, *computer* collocates with *instructions*.

4.1.6 Discussion

Level of keyness is the first criterion to be taken for the presence of metaphor. Secondly, the frequency and its percentage of lemmas need to be taken into account as well, whether of key words or non-key words. In Table 1 we can see that the key words of the *Office* category are the highest in keyness with four constituents in the first 100, and the fourth in the first 200. The most comprehensive one is the *Workshop* metaphor. However, it has only one constituent in the first 100. It has three constituents in the first 200 and five in the first 300. *Living Being* metaphor comes third. It has five key words with two in the first 100, two in the first 400 and one is the last in the list. The *Soldier* metaphor has two key words only with one in the first 50 and the second in the first 150. The *Place/Building* metaphor occupies the bottom of the list with two key words in the first part of the first 200.

Out of the total key words of this category, the *Office* metaphor is the highest (43.51%), followed by the *Workshop* metaphor (23.85%). At the same time, the *Office* metaphor has the highest percentage of the whole corpus (1.2%) in comparison to other metaphors in this category. The computer metaphors constitute 3% of the whole corpus.

	Living Being	Workshop		Office	Place/Building	Soldier	Total
Order of KWs	26 64 365 345 500	46 113 155 194 218 220	230 269 296 354 425 457	2 28 34 68 173	104 148	39 146	
Frequency of KWs	20855	32342		58997	9875	13523	135592
Percentage of this category KWs to all *Computer* KWs	15.38	23.85		43.51	7.28	9.97	
KW Lemmas	25471	55039		82045	10698	13812	
Frequency of non-KW Lemmas	12561	5901		4098	5505	1277	
Total of Lemmas	38032	60940		86143	16203	15089	216407
Percentage of these lemmas to the whole Corpus	0.53	0.85		1.2	0.22	0.21	3

Table 1: Statistics of the *computer* metaphors

4.2 The Internet

Internet is a key word (38) ranking much higher than the *computer* (129). *Internet* collocates with most of the category constituents. Another name that is used for *Internet* is *web*. We will also look at this word.

4.2.1 THE INTERNET IS IN A STATE OF WAR

In this category we have two key words: *password* (84) and *security* (327). Their percentage to the total frequency of key words in the Internet metaphor categories is 13.34%. By including lemmas of non-key words such as *war*, *crack* and *bomb*, the percentage of the constituents of this category is 0.47% of the whole corpus. *Internet* collocates with *warfare, security* and *cracking*. *Bombed* occurs three times with *Internet* as its object in one of them. *Web* was found to collocate with *secure*.

4.2.2 THE INTERNET IS A ROAD

We have one key word *path* (346). Its percentage to the total frequency of key words in the Internet metaphor categories is 4.3%. By including lemmas of non-key words such as *road*, *traffic*, *highway*, *bus*, and *map*, the percentage of the constituents of this category is 0.06% of the whole corpus. *Internet* collocates with *traffic, shortcut, speed* and *transport*. *Web* was found to collocate with *traffic*. On the other hand, we have collocations such as *data transport*, *data highway*, and *information superhighway* that imply the metaphor.

4.2.3 THE INTERNET IS A BUILDING/PLACE:

Here we have two key words: *access* (32), *address* (157) and *site* (426). Their percentage to the total frequency of key words in the Internet metaphor categories is 36%. By including lemmas of non-key words such as, *visit*, *go*, *firewall*, *architecture*, *portal*, *gateway*, *home* and *wallpaper*, the percentage of the constituents of this category is 0.49% of the whole corpus.

Collocations are found to support this category. *Internet* collocates with *access*, *address*, *local*, *location* and *site*. A top collocate of *site* is *web*. *Visit* has *web*, *site* and *internet* as well as many *URLs* as object collocates. When checking the collocation pattern of *http*, which is a part of internet addresses, it was found collocating with location words such as *here*, *located at*, *available at,* and *found at*. *URL* has *address* and *destination* as top collocates. Here are some examples of the verb collocates *go* and *visit*:

Click the http://www.3com.com to go to 3Com's World Wide Web site.

Go to any website anywhere…

… the objects you encounter as you visit Internet sites…

Internet collocates with *access* in the L1 slot. All top L1 verb collocates of *access* imply permission: *gain*, *grant*, *delegate*, *restrict*, *allow*, *control*, *prevent*, and *provide Unauthorized* is a top L1 collocate as well.

4.2.4 THE INTERNET IS A BOOK

Key words include *page* (19), *pages* (87), *browser* (100), *browse* (429) and *bookmark* (471). Their percentage to the total frequency of key words in the Internet metaphor categories is 46.3%. By including lemmas of the non-key word *publish*, the percentage of the constituents of this category is 0.43% of the whole corpus.

Page and its plural form have both *web* and *Internet* as collocates in the L1 slot, though the collocation is much more frequent with the first. *Browse* has both *web* and *Internet* as object collocates. Both *web* and *Internet* co-occur with *browser* in the L1 slot. Other collocations that were found to support this category are: *Internet publishing*, *publishing Web pages*, *To publish Web pages* and *Republish web pages*.

4.2.5 THE INTERNET IS A SEA

No constituent of this category was found to be a key word. *Navigator* was found to be a key word, but has been excluded because it is a part of a brand name of the program *Netscape Navigator*. However, the name implies the metaphor. Non-key words are *navigate*, *pirate* and *surf*. These make 0.014% of the whole corpus. *Internet* was found to be an object collocate of *navigate*. One collocation that was found to imply the metaphor is *data stream*.

4.2.6 THE INTERNET IS A MARKETPLACE

For this category no key words was found. Percentage of non-key word lemmas to the whole corpus is 0.01%. *Internet* collocates with *marketing*, *commerce*, *e-commerce* and *e-marketing*.

4.2.7 Discussion

To summarize the statistics of the Internet metaphors, we can see in the table below that the *Book* metaphor is the most prominent one in terms of keyness followed by the *Building/Place* metaphor. In terms of percentage of the corpus, the *Building/Place* metaphor is the highest followed by the *War* metaphor and the *Book* metaphor. The *Marketplace*, *Sea* and *Road* metaphors are marginal although the latter has one key word. The first two has no key words.

	State of War	**Road**	**Building/Place**	**Book**	**Sea**	**Marketplace**	**Total**
Order of KWs	84 327	346	32 157 426	19 87 100 429 471	----	----	
Frequency of KWs	8591	2774	23203	29821	-----	----	64389
Percentage of this category KWs to all *Internet* KWs	13.34	4.3	36	46.3	----	----	
KW Lemmas	32972	3275	26874	30782	-----	----	
Frequency of non-KW Lemmas	1326	1342	8644	731	1009	756	
Total of Lemmas	34298	4617	35518	31513	1009	756	107711
Percentage of these lemmas to the whole Corpus	0.47	0.06	0.49	0.43	0.014	0.01	1.5

Table 2: Statistics of the Internet metaphors

The figures in tables 1 and 2 show that the *computer* metaphors are more common in LIT than *Internet* metaphors. Key words of the *computer* metaphors make 67.8% of the total frequency of key words of both the *computer* and the *Internet* metaphors, whereas those of the *Internet* metaphors are

32.19%. On the other hand, both categories of the *computer* metaphors and the *Internet* metaphors make 4.5% of the whole corpus, of which the *computer* metaphors make 3% and the *Internet* metaphors make only 1.5%. The *Office* metaphor is the highest in the whole corpus.

5. Conclusion

In using corpora in the study of figurative language, key word and frequency lists help in mapping out the use of metaphor, especially in a special variety of language. This has to be based on our conceptual perspective of the language use. Hence an interaction between the two approaches is important to have an overview of how the figurative use is organised. However, the same words can be found in the reference corpus in their literal meaning and thus affect the level of keyness negatively. On the other hand, results from a corpus-based study help in adjusting our conceptual metaphors or adding constituents to the categories, e.g. *sleep* which was not included in the initial stage of research.

Metaphor is highly used in LIT. Having key words of a metaphoric profile is an evidence from a fairly large corpus that metaphor is persistent in IT. At the same time the principal elements of IT, i. e. *computer* and *Internet* have metaphorical collocates. Collocation is an important indicator when key words collocate with words that denote the category or imply the metaphor.

References

Apple 1987 *Human interface guidelines: the Apple desktop interface*. Reading, Massachusetts, Addison–Wesley.

Beck C 1991 Implications of metaphors in defining technical communication. *Technical Writing and Communication* 21(1) 3-15.

Black M 1962 *Models and metaphors*: *studies in language and philosophy*. Ithaca, London, Cornell University Press.

Chisholm R 1986 Selecting metaphoric terminology for the computer industry. *Technical Writing and Communication* 16(3) 195 -220.

Cohen L 1993 The Semantics of metaphor. In Ortony A (ed.) *Metaphor and thought*, 2nd edition, Cambridge, Cambridge University Press. pp 58 -70.

Constantine L 2001 *The peopleware papers*. NJ, Yourdon Press.

Coyne R 1995 *Designing information technology in the postmodern age, from method to metaphor*. Cambridge, Mass. and London, MIT.

Goatly A 1993 Species of metaphor in written and spoken varieties. In Ghadessy M (ed.), *Register analysis: theory and practice*. London and NY, Printer. pp 110 -148.

Goatly A 1997 *The language of metaphors*, London, Routledge.

Grevy C 1999 Informationsmotorvejen og andre metaforer i computerfagsprog. *Hermes* 23: 173 -201.

Gross A 1990 *The rhetoric of science*. Cambridge, Harvard University Press.

Hesse M 1980 *Revolutions and reconstructions in the philosophy of science*. Bloomington, Indiana University Press.

Izwaini S 2000 *Translating collocation: Arabic/English/Swedish*. Unpublished MSc dissertation. University of Manchester Institute of Science and Technology (UMIST).

Izwaini S 2003 Building specialised corpora for translation studies. In *Proceedings of the Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives, 27 March 2003, Corpus Linguistics 2003,* Lancaster. pp 17 -25.

Johnson G 1991 Agents, engines, traffic, objects and illusions; paradigms of computer science. *Technical Writing and Communication* 21(3) 271 -283.

Kittay E, Lehrer A 1981, Semantic fields and the structure of metaphor. *Studies in Language* 5 (1)31 - 63.

Lakoff G 1992 Metaphor and semantics. In Crystal D (ed.), *International encyclopedia of linguistics*. Oxford, OUP.

Lakoff G 1993 The contemporary theory of metaphor. In Ortony A (ed.), *Metaphor and thought*, 2nd edition, Cambridge, Cambridge University Press. pp 202-251.

Lakoff G, Johnson M 1980 *Metaphors we live by*. Chicago, University of Chicago Press.

Levin S 1977 *The semantics of metaphor*. Baltimore and London, John Hopkins University Press.

Martin J 1996, Computational approaches to figurative language. *Metaphor and Symbolic Activity* 11(1): 85-100.

Meyer I, Zaluski V, Mackintosh K 1997 Metaphorical internet terms: a conceptual and structural analysis. *Terminology* 4(1): 1 -33.

Microsoft, 1993 *The GUI guide: international terminology for the windows interface*. Redmond, Washington.

Mulder M 1996 Perception of anthropomorphistic expressions in software manuals. *Technical Writing and Communication* 26(4) 489 -506.
Ricoeur P 1978 *The rule of metaphor: multi-disciplinary studies of the creation of meaning*. London, Routledge and Kegan Paul.
Rothbart D 1984 The semantic of metaphor and the structure of science. *Philosophy of Science* 51: 595 -615.
Scott, M., 1997, *WordSmith tools, version 2*.0, Oxford, Oxford University Press.
Searle J 1993 Metaphor. In Ortony A (ed.), *Metaphor and thought*, 2nd edition, Cambridge, Cambridge University Press. pp 83-111.
Öberg P 1989 Metaphors we compute by. In Odenstedt B, Persson G (eds), *Instead of flowers: papers in honour of Mats Rydén on the occasion of his sixtieth birthday*. Stockholm, Almqvist and Wiksell.

Interdisciplinary Workshop on Corpus-Based Approaches to Figurative Language. 27 March 2003, Lancaster University

Metaphor corpora and *corporeal* metaphors

Andreas Musolff (University of Durham)

1) Introduction

Cognitive metaphor theory has highlighted the conceptual function of metaphor by providing evidence of domain-mapping systems that make up our universe of experiences. Is this conceptual function also of relevance in the public political debate? Lakoff and Johnson's answer is an emphatic *Yes*: "Metaphors may create realities for us, especially social realities. A metaphor may thus be a guide for future action. Such actions will, of course, fit the metaphor" (1980: 156). Lakoff himself and others have since produced a number of empirical case-studies of the role metaphors play in public discourse in national as well as international politics (e.g. Chilton and Lakoff 1995; Lakoff 1996, 2001; Schäffner 1996; Dirven, Frank and Ilie 2001). In many of these studies, however, we find a tension between, on the one hand, strong general claims such as the one quoted above which suggest that metaphor sources 'guide' social and political practice up to the point of acting as self-fulfilling prophecies and, on the other hand, actual empirical findings that reveal a "wide variety of possible entailments" of one source domain, which offer "scope for debate and controversy" (Schäffner 1996: 56). If the same source domain can be used to argue for or against specific political positions, its 'guiding' force evidently is ambiguous. It is here that corpus-based studies are needed because they allow us to go beyond illustrating cognitive hypotheses with a few 'fitting' examples by way of eliciting empirical data on whether are politically significant differences in the ways metaphor source domains are used in particular discourse communities.

The basis for the presentation is a bilingual corpus of texts containing metaphorical references to European politics in the 1990s has been built up, drawn from British and German press coverage of political decisions and developments concerning the European Union. The corpus has a pilot version, called EUROMETA I, which comprises some 2100 passages from 28 British and German newspapers from the period 1989-2000, and a larger version (EUROMETA II) compiled from two general corpora, i.e. the "Bank of English" at the University of Birmingham and "COSMAS" at the Institute for German Language in Mannheim, which comprises in excess of 20.000 entries. For the present study, metaphors from the source domain of LIFE-BODY-HEALTH have been selected. This source domain is among the most fundamental and ancient metaphor systems employed for the conceptualisation of socio-political entities, reaching back, in the Western tradition, to ancient and medieval concepts of the state as a *body politic* (Hale 1971, Struve 1978, Sontag 1991). A first overview shows that by no means all aspects of the source domain BODY-HEALTH-ILLNESS are employed in the modern Euro-political debates, and secondly, that there frequency of use differs greatly, as can be seen from tables 1 and 2.

Table 1) Conceptual elements of the LIFE-BODY-HEALTH domain in EUROMETA II

Source concepts	English lexemes	German lexemes
LIFE-SURVIVAL	life, alive, live, survival	Leben, leben, lebendig, über-, weiterleben, ins Leben rufen
BIRTH-BABY	birth, rebirth, born, still-born, premature birth, abortion, baptism, baby, (bouncing) child,	Geburt, geboren, Wiedergeburt, Frühgeburt, Missgeburt, Kind, Baby
DEATH	death sentence/ warrant/ knell	Tod, tot
ILLNESS/DISEASE		
I/D: SICK/ILL	*Ill, illness, sick (sick man of Europe)*	*krank, kranker Mann Europas, kränkelnd*
I/D: EUROSCLEROSIS	*Euro(-)sclerosis*	*Eurosklerose*
I/D: MADNESS	*(Euro-)madness*	
I/D: INFLUENZA	*Asian (economic) flu*	Grippe
I/D: VIRUS	*virus*	
I/D: COLIC	*colic*	
I/D: WOUND		*Wunde, Narbe*
I/D: WASTING/TBC		*Schwindsucht*
I/D: HURT		*wehtun*
CURE/THERAPY/CARE	therapy, diagnose	Pflege, pflegen, Nachsorge
HEALTH/FITNESS/ RECOVERY	recovery, health, healthy	Gesundheit, gesund, gesünder, gesunden (v.), Fit, Fitness, Erholen
BODY PARTS		
BP: HEART	*heart*	*Herz*
BP: EYES		*Augen*
BP: HEAD		*Kopf*
BP: LEGS		*Beine*
BP: FEET		*Füße*
BP: MUSCLES		*Muskeln*
BP: BACKSIDE	*backside*	

* Not including lexicalised imagery for political leaders as *heads of state/government/commission* (cf. Deignan 1995: 1-2).

Table 2) Tokens for conceptual elements of LIFE-BODY-HEALTH source concepts in EUROMETA II in the order of overall frequencies

Source concepts	**number of tokens in English sample**	*	**number of tokens in German sample**	*	**overall number of passages**
BODY PARTS	210		377		587
BP: HEART		*209*		*336*	
BP: EYES				*19*	
BP: HEAD				*9*	
BP: LEGS				*6*	
BP: FEET				*5*	
BP: MUSCLES				*2*	
BP: BACKSIDE		*1*			
ILLNESS/DISEASE	60		137		197
I/D: SICK/ILL		*40*		*92*	
I/D: EUROSCLEROSIS		*12*		*32*	
I/D: MADNESS		*4*			
I/D: INFLUENZA		*2*		*3*	
I/D: VIRUS		*1*			
I/D: COLIC		*1*			
I/D: WOUND				*5*	
I/D: WASTING/TBC				*3*	
H/I: HURT				*2*	
BIRTH-BABY	58		100		158
HEALTH/FITNESS/ RECOVERY	37		111		148
LIFE-SURVIVAL	23		55		78
DEATH	4		8		12
CURE/THERAPY/ CARE	2		7		9

TOTAL	**394**		**795**		**1189**
no. of passages	184		485		669

* figures for tokens in columns 3 and 5 are included in the figures of the preceding columns and thus do not add to total sums

The overall ratios of German and British sample passages and tokens (2.6:1 and 2:1) should not be seen as evidence of greater a popularity of LIFE-BODY-HEALTH metaphors in German press language but are due to the fact that the German corpus contains many more texts for the same period (1989-2001) than the BoE. It would thus be misleading to derive conclusions from the absolute frequency of lexical items or source elements. The following remarks will instead focus on differences between in distribution patterns the two national samples. To do this, I have grouped the concepts into "scenarios" (Lakoff 1987: 285-286) based on three central mappings of LIFE-BODY-HEALTH sources to the target domain of political INSTITUTIONS:

(1) AN INSTITUTION HAS A **life cycle** that lasts from birth to death

scenarios: AN INSTITUTION IS CONCEIVED, CARRIED AND BORN; IF IT CONTINUES TO FUNCTION IT SURVIVES AND GROWS UP; WHEN IT CEASES FUNCTIONING, IT DIES

(2) AN INSTITUTION CAN BE IN A MORE OR LESS **healthy/ill** state

scenarios: THE INSTITUTION CAN SUFFER INJURIES OR FALL ILL, RECOVER, AND UNDERGO MEDICAL TREATMENT

(3) AN INSTITUTION HAS A **body** that comprises various parts

scenarios: THE PARTS OR ASPECTS OF AN INSTITUTION ARE LIMBS AND ORGANS OF ITS BODY (which can also individually become ILL – see (2) - and then may affect the whole BODY).

2) The *life cycle* of Europe

The main focus of the LIFE-CYCLE scenario in both EUROMETA corpora is the concept of BIRTH. It is used to describe momentous and innovative political developments, such as the restructuring of Europe after the collapse of the Warsaw Pact in 1989/90, or the institutional reforms of the EC/EU which were agreed in the Treaties of Maastricht (1991) and Amsterdam (1997). More than 75% of all BIRTH tokens are, however, references to the common currency, the "euro". They comprise a variety of pre-, peri- and post-natal problems as well as emphatically positive descriptions of a HEALTHY BABY. When we look at the distribution of these scenario versions, the tokens for the characterisation of the euro introduction as a PROBLEM BIRTH in the British sample account for less than 40 % (9 out of 26 in EUROMETA II) but they make up the great majority, 63%, in the German sample (50 out of 79 in EUROMETA II). A closer study of the reasons for this apparent lack of confidence in the *euro-birth* in the German public reveals an interesting characteristic of the German EU-debate as well as a specific problem for the interpretation of statistical corpus data. 90% of all the German tokens consist of citations of and comments on a particular statement by one politician, namely the then opposition contender for the German Chancellorship, the Social Democrat Gerhard Schröder. In a statement made at a crucial time in the run-up to the federal elections in Germany, in March 1998, Schröder warned that a hastened arrival of the euro

was going to *deliver a sickly, premature baby* ("eine kränkelnde Frühgeburt"; W, 27/3/1998). The incumbent Chancellor, Helmut Kohl, and his foreign minister, Klaus Kinkel, launched a counter-attack by condemning Schröder for having denounced the euro as a case of *miscarriage* or even monstrosity (i.e. a "Fehlgeburt" or "Missgeburt"; SZ, 28/3/1998, MM, 3/4/1998). Schröder hit back by accusing his opponents of misquoting him and demonstrated how carefully he had chosen his metaphor scenario: of course, he stated, 'a *miscarriage* and a *premature birth* were completely different things' (SZ, 28/3/1998). Calling the *birth* of the euro a *miscarriage* would have implied an utterly pessimistic attitude, namely, that the *child was doomed to die*, whereas the diagnosis of a *premature birth* could be interpreted as a plea for *extra care* and *support* so that *the child could still survive*. Needless to say, Schröder considered his own party to be *best qualified to give that support* and he reiterated this assessment on the election campaign trail (SP, 14/1998 = example 2). After winning the elections, Schröder was charged, as acting president of the EU Council of ministers from January to June 1999, with *caring* for the *child* whose allegedly *premature birth* he had criticised only a few months before. In case he might have forgotten his earlier words, the magazine *Der Spiegel* (SP, 1/1999 = example 3) was quick to remind him of that somewhat premature diagnosis. As late as May 2000, the German press harked back to Schröder's 1998 statement, using it as a foil for evaluating his arguments that the (then topical) decline of the euro's in the exchange rate against the US dollar was 'nothing to shed tears over' (MM, 12/5/2000).

Such debates had a massive impact on the distribution patterns of BIRTH metaphors in the corpus. From the start of the 1990s until spring 1998, BIRTH metaphors occur on average once a year. PROBLEM BIRTH scenarios then begin to pick up in both samples, which can be explained with reference to the approaching date for the currency introduction. This increase is, however, insignificant compared with the sudden inflation of tokens for the PREMATURE BIRTH scenario, which dominate the German sample and make up the bulk of all BIRTH tokens up to October 1998 (i.e., the time of the general election). After that their frequency decreases but still remains at a higher level than before March 1998. We can thus observe how a sub-group of scenarios focussing on a special target topic provide the bulk of tokens in the corpus, due to developments that have nothing to do with the conceptual or ontological centrality of the scenario but rather with social and political dynamics. Once introduced in the public debate by a prominent politician in a salient context, a scenario is disseminated through quotations and comments in the media. In the short term, the metaphor remains the property, as it were, of its author and serves his argumentative needs – thus, the PREMATURE BIRTH metaphor helped Schröder to sound sufficiently Euro-sceptical and to ingratiate himself with the German electorate of 1998. Within a changed political context, however, the metaphor was quoted against him in the context of comments suggesting that Schröder's *premature birth* warnings had sounded too sceptical to be offset easily by his new posturing as a *caring euro father*. The lesson for the interpretation of corpus data is that the frequency of occurrences of tokens for conceptual elements or scenarios cannot in itself be regarded as evidence of an argumentative or ideological bias of the source as used in a given discourse community.

3) *Health* and *illness* of Europe

Whilst the PREMATURE BIRTH scenario includes SICKNESS only in some variants, there are also scenario versions in which the concepts of ILLNESS, TREATMENT and RECOVERY are explicitly foregrounded, comprising

physical, psychological and psychosomatic conditions and concomitant effects such as PAIN and AGONY, as well as forms of PREVENTION and THERAPY, e.g. FITNESS PROGRAMMES, HEALTH TESTS, PILLS OR MEDICINES. General terms for concepts in the ILLNESS domain, namely English *disease*, *illness*, *ill* and *sick* and German *Krankheit*, *krank* as well as *kränkeln* ('being poorly', 'beginning to be ill') collocate in some cases with specific LIFE/ILLNESS terms, but most of their tokens appear in the fixed phrase *the sick man of Europe* (in German: *der kranke Mann Europas*). This phrase is by no means new to European discourse – the *sick man of Europe* formula can be traced back to the late 17th century but gained prominence in the 19th century as a reference to the declining military and economic power of the Ottoman Empire (Büchmann: 1898: 531-514; *Brewer's* 2001: 1083-1084). More recently, in the 1970s, Britain had also been dubbed the *sick man of Europe*. It was with considerable relief, and in a few cases with *schadenfreude*, that British media passed the stigma label onto Germany, when the erstwhile model of a *healthy economy* in Europe experienced the double threat of recession and of not meeting the EMU stability criteria (T, 26/10/2001). Altogether, the *sick man* references make up 29 of the 40 British SICK/ILL(NESS) tokens in EUROMETA II, i.e. 72%., with most of them harking back to Britain's past status as the *sick man* (15 tokens). Germany comes second (10 tokens); apart from it only Albania and the euro carry this stigma (each just once). On the German side, the *sick man of Europe* theme is much less prominent: there are overall just 13 tokens of *kranker Mann Europas* (= 14% out of 92), and Germany is by no means the only target: there are also tokens for Russia, Greece and Spain, plus acknowledgements that Britain has left the negative image behind.

Whilst the *sick man of Europe* phrase presupposes a mapping from the concept of SEVERAL PERSONS to the DIVERS EUROPEAN STATES, so that one of them can be identified as *the* SICK MAN, an alternative perspective is that of the EU as one integral ORGANISM THAT SUFFERS FROM A ILLNESS OR DISEASE. This concept underlies the two remaining ILLNESS scenarios manifested in the EUROMETA corpora, i.e. the scenario of a special ILLNESS called *Euro-sclerosis* and that of an affliction of THE HEART OF EUROPE. *Euro-sclerosis* is the only medical metaphor in Euro-discourse that has achieved the status of a well-known key-term that can be identified 'in its own right', as it were, in a general corpus. It seems to have been coined first in the 1980s to warn against a decrease in economic and institutional *flexibility* and *growth* (Jung and Wengeler 1995: 110), in keeping with its source meaning of "morbid tissue hardening" (*The Concise Oxford Dictionary* 1979: 1014-1015). During the second half of the 1990s, the verdict of *Euro-sclerosis* was used predominantly in the British press to condemn low growth and rising unemployment in continental member states and to promote abstention from Monetary Union (T, 25/12/2000).

4) The *heart* of Europe

Of the altogether nine body parts named in texts of the two EUROMETA corpora, only one constitutes a significant source concept for Euro-metaphors: i.e. the HEART. The remaining BODY PART concepts appear, in special one-off formulations that have the HEART source concept as their principal semantic clue. Two main types of folk-theories seem to be activated in the analogy *HEART:BODY* to *X:EUROPE*, as documented in the corpus: a) an understanding of the HEART as the CENTRAL PART of the BODY, and b) the notion that the HEART as an ORGAN can suffer damage from INJURY or DISEASE.

The CENTRALITY aspect of the HEART concept serves, first of all, as a reference to countries, regions or

cities as being situated geographically at the *heart of Europe*. They are statistically by far the most prominent uses of the phrase *heart of Europe* in the EUROMETA II German sample (with 257 out of 336 tokens), and still make up a sizeable portion in the English sample (34 out of 209). Nearly half (i.e. 116 out of the 252) German tokens relate to Germany as a whole or German cities as being the *heart of Europe* or as being *in the heart of Europe.* There are no similar references to Britain - in either the British or the German sample of the corpus. As regards continental Europe, the HEART=CENTRE equation extends not just over the countries of central Europe - i.e., Poland, the Czech Republic, Germany, Austria, Slovenia - but also includes Belgium, the Franco-German border regions (Alsace-Lorraine, Burgundy and the Palatinate), Switzerland, and the Balkans. The latter occur mainly in references to the wars in the former Yugoslavia as taking place *in the heart of Europe*, with the implication that what happens *in the heart* is – or should be – close to, and of particular importance for, one's emotional centre (cf. (G, 5/4/1999). This emotive dimension of positioning a nation *in the heart of Europe* is also discernible in references to candidate states for the EU enlargement process, such as the Czech Republic, Poland and Hungary (taz, 2/1/1995). The localisation of a nation (or metonymically, its capital) *in the heart of Europe* carries with it the demand or promise that it has a right to be taken seriously as a member of a united Europe.

This implicit bias is even stronger when we move on to non-geographical *heart of Europe* categorisations. In this context, Britain finally comes into the picture – indeed, the British public debate about EC/EU-politics in the 1990s can in some sense be summarised as a dispute about the nature and function of the *heart of Europe* and Britain's relation to it. At the beginning of this debate stands again another prominent key-utterance by John Major, in a speech held four months after he had succeeded Margaret Thatcher as British Prime Minister and Tory Party leader. In it he pledged that "[...] Britain would work 'at the very heart of Europe' with its partners in forging an integrated European community" (G, 12/3/1991). Over the following months, Major's *heart of Europe* slogan triggered a host of interpretations and variations. For a while, the majority of interpretations and comments were consistent as regards the scenario of WORKING AT THE HEART OF AN INSTITUTION, i.e. they treated it as equivalent to the notion of BEING CLOSELY INVOLVED WITH IT. In 1994, the joint parliamentary groups of the ruling German Christian Democrat parties even used the reference to Major's statement in a manifesto to express their 'hope that 'Britain should play its role at the heart – i.e., at the core - of Europe'. The *Guardian* commented that the paper was "by far the most important recognition by a political body indisputably - as opposed to rhetorically - at the heart of Europe that the Maastricht project will now be rethought" (G, 3/9/1994). The thinly disguised condemnation of the Conservatives as being 'only rhetorically at the heart of Europe' signalled that the dominant political interpretation of the BRITAIN-AT-THE-HEART-OF-EUROPE notion had changed. Major's claim from 1991 was by now seen as hollow. In the following years German media repeatedly quoted Major's promise as evidence against his apparent turn-around to a Euro-sceptical position, in way not unlike their strategy of confronting Chancellor Schröder with his former *premature birth* misgivings. In the British debate, similar attempts were made to remind Major of his erstwhile Euro-enthusiasm but there were also more direct challenges to his BRITAIN-AT-THE-HEART-OF-EUROPE promise. With the integration process slowing down after several delays in the ratification of the Maastricht Treaty and the withdrawal of the Pound Sterling from the European Exchange Rate Mechanism, the originally intended positive appeal to CENTRAL involvement in the EU as expressed in the *heart* metaphor lost some of its plausibility, and Major's

phrase was adapted to pessimistic scenarios of an imminent HEART FAILURE (I, 11/9/1994). As if this was not enough to give the *heart of Europe* a bad name, a further scenario emerged with the publication of a strongly Euro-critical book written by the EU official Bernard Connolly in 1995, entitled *The Rotten Heart of Europe.* It captured the headlines of the British press (e.g. E, 9/9/1995; G, 11/9/1995), as well as leading to his sacking by the EU commission. The ROTTEN HEART scenario constitutes a special blending, insofar as the well-established mapping ENTITIES THAT ARE DETERIORATING ARE ROTTEN or ROTTING ORGANISMS, is applied to the concept of HEART in its metaphorical meanings of CENTRE and CHIEF ORGAN of the EU. It thus conveys a sense of a particularly dangerous type of deterioration which is hard to heal, if at all.

By contrast with the British debate, SICK, ILL or ROTTEN HEART metaphors are rare in the German sample: there are only two occurrences. The remainder are neutral or positive, and a substantial sub-section, i.e. 25 tokens, is made up of references to the British debate. On the occasion of the 1999 nepotism scandal we find - despite massive critical coverage -, no HEART FAILURE imagery in the German media but just one mildly ironical reference by the *Frankfurter Rundschau* to Blair as a would-be '*dragon slayer at the heart of Europe*' (FR, 24/3/1999).

Such use of the phrase *putting Britain at the heart of Europe* with regard to Blair betrayed a good knowledge of British debates. Blair had 'inherited', as it were, the role of promoter of the *Britain at the heart of Europe* slogan from his predecessor. Together with that claim he has inherited the challenges to it in the form of DISEASE/ILLNESS scenarios: at the end of Blair's first term of office, a *Guardian* article depicted him as a man, who talks about being *at the heart of Europe*, but when arrives there might be received as someone in need of " *a look of pity and a cup of sweetened tea — but only after he has wiped his feet in a trough of disinfectant*." (G, 4/4/2001). This example links the slogan of *being at the heart of Europe* with an allusion to the then topical "foot-and-mouth" epidemic in Britain. Even though there is no connection between the HEART concept and the EPIDEMIC scenario, the latter provides the thematic perspective of HEALTH/HYGIENE PROBLEMS that affects the understanding of *heart of Europe*, suggesting inferences that the EU might not want Britain to be *close to its heart* because of its perceived *sickness*. This produces a sarcastic effect of exposing a perceived lack of realism and common sense among the *heart of Europe* supporters.

5) Conclusions

This survey of LIFE-HEALTH-BODY metaphors from the EUROMETA corpora demonstrates that elements of conceptual source domains and their configurations in scenarios can be found in a general corpus by defining using key-word searches consisting of pairs of lexical items that belong to the source and target domains. On the other hand, the analysis shows that conceptual domains as such cannot provide a sufficiently well-defined basis to explain the *distribution* patterns of source concepts that are characteristic for the respective discourse communities. The evidence from the LIFE-HEALTH-BODY domain data leads to the hypothesis that within a domain certain elements and scenarios have a privileged, prominent status in that they account for most of the metaphor tokens as well as for their textually most elaborate variations. In the course of public debates within a discourse community, traditions of metaphor use emerge in which specific scenarios (e.g. *premature birth, being at the heart of Europe, Euro-sclerosis, the sick man of Europe* etc.) become the foci of extensions and re-interpretations and 'conceptual contests' – hence a sudden inflation of tokens for the respective scenarios in the cor-

pus at particular points in the communicative history of that community. Some of these contests become so prominent that they are reported in a neighbouring discourse community (e.g. British claims of being *at the heart of Europe* that were commented on in German media). In the course of these debates, the ideological and argumentative bias of source concepts may change drastically. Thus, the initially optimistic-sounding phrase *being close to the heart of Europe* was turned against its authors in comments that highlighted *diseases* of or *injuries* to that *heart*; and Schröder's verdict on the euro's *birth problems* was quoted against him as well as being twisted around by himself. By focusing on such traditions of usage, corpus-based analysis can highlight argumentative tendencies and ideological assumptions that are associated with specific scenarios rather than with the abstract level of source domains. The grouping of lexical and phraseological items in specific scenarios and the distribution of scenarios in a public discourse corpus can thus be understood as indicators of thematic and argumentative perspectives that are representative for a discourse community.

References

Altenberg, Bengt and Granger, Sylviane 2002. *Lexis in Contrast. Corpus-based approaches*. Amsterdam/Philadelphia: Benjamins.

Brewer's Dictionary of Phrase and Fable 2001. Ed. Adrian Room. London: Cassell.

Büchmann, Georg 1898. *Geflügelte Worte*. K. Weidling (ed.). Berlin: Haude and Spener'sche Buchhandlung.

Chilton, Paul and George Lakoff 1995. Foreign Policy by Metaphor. In: Christina Schäffner and Anita Wenden (eds.). *Language and Peace.* Dartmouth: Aldershot, 37-59.

Concise Oxford Dictionary 1979. J. B. Sykes (ed.). Oxford: Oxford University Press.

Deignan, Alice 1995. *COBUILD English Guides. 7: Metaphor Dictionary* London: HarperCollins.

Deignan, Alice 1999. Corpus-based research into metaphor. In: Lynne Cameron and Graham Low (eds.). *Researching and Applying Metaphor*. Cambridge: Cambridge University Press, 177-199.

Dirven, René, Roslyn M. Frank & Cornelia Ilie 2001 (eds.). *Language and Ideology. Volume II: Descriptive Cognitive Approaches*. Amsterdam/Philadelphia: John Benjamins.

EUROMETA-corpus 2002. http://www.dur.ac.uk/SMEL/depts/german/Arcindex/htm.

Hale, David 1971. *The Body Politic. A Political Metaphor in Renaissance English Literature*. The Hague: Mouton.

Hunston, Susan 2002: *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Jung, Matthias and Martin Wengeler 1995. *Nation Europa* und *Europa der Nationen*. In: Georg Stötzel and Martin Wengeler (eds.). *Kontroverse Begriffe. Geschichte des öffentlichen Sprachgebrauchs in der Bundesrepublik Deutschland*, Berlin/New York: de Gruyter, 93-128..

Lakoff, George 1987. *Women, Fire, and Dangerous Things. What Categories Reveal about the Mind*. Chicago/London: University of Chicago Press.

Lakoff, George 1996. *Moral Politics: What Conservatives Know That Liberals Don't*. Chicago/London: University of Chicago Press.

Lakoff, George 2001. September 11, 2001. *Metaphorik.de*: http://www.metaphorik.de/aufsaetze/lakoff-september11.htm.

Lakoff, George and Mark Johnson 1980. *Metaphors we live by.* Chicago: University of Chicago Press.

Musolff, Andreas 2000. *Mirror Images of Europe. Metaphors in the public debate about Europe in Britain and Germany.* Munich: iudicium.

Musolff, Andreas 2001. Cross-language metaphors: *parents* and *children, love, marriage* and *divorce* in the *European family.* In: Janet Cotterill and Anne Ife (eds.). *Language across Boundaries*. London/New York: Continuum, 119-134.

Niemeier, Susanne 2000. Straight from the heart — metonymic and metaphorical explorations. In: Antonio Barcelona (ed.). *Metaphor and Metonymy at the Crossroads. A Cognitive Perspective*. Berlin/New York: De Gruyter, 195-213.

Röhrich, Lutz 2001. *Das große Lexikon der sprichwörtlichen Redensarten*. (Neuausgabe). 3 vol.s Darmstadt: Wissenschaftliche Buchgesellschaft.

Schäffner, Christina 1996. Building a European House? Or at Two Speeds into a Dead End? Metaphors in the Debate on the United Europe. In: Andreas Musolff, Christina Schäffner and Michael Townson (eds.). *Conceiving of Europe — Unity in Diversity.* Aldershot: Dartmouth, 31-59.

Sontag, Susan 1991. *Illness as Metaphor. Aids and its Metaphors*. Harmondsworth: Penguin.

Struve, Tilman 1978. *Die Entwicklung der organologischen Staatsauffassung im Mittelalter.* Stuttgart: Anton Hiersemann.

Using LSA to detect Irony

Aynat Rubinstein, Department of Linguistics, Tel Aviv University

Abstract

In this work I propose a new model of verbal irony based on the notion of scales. The model, which stems from discourse theoretic accounts for irony, is then given computational concreteness based on Latent Semantic Analysis (LSA) [1]. Preliminary results are presented for automatically detecting irony in ironic headlines, a special type of irony which we argue is most fit for the LSA analysis.

Irony on a Scale

From a discourse theoretic perspective, the model of scales suggests that understanding irony means perceiving the distance between two points on a scale [2].

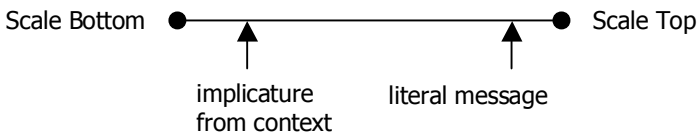


Figure 1: A scale for irony

A scale is the structure depicted in Figure 1 above. It is a line representing degrees, or cases, of the discourse topic alluded to in the utterance. The discourse topic defines the content of the scale and its edges, which represent extreme opposite cases of its realization. In understanding an ironic utterance, one point is conveyed by the literal meaning of the utterance, and the other is a relevant implicature extracted from context. The greater the difference between the two points, it is claimed, the better the resulting irony in terms of ease of perception and appropriateness.

For example, consider the following situation: your parents are away for the weekend, the house is totally at your disposal, it is Saturday afternoon and you have invited your boyfriend over. Just as the two of you are getting intimate on the sofa, your parents suddenly walk in. "What perfect timing!", you exclaim when you see them. Your boyfriend probably understands your ironic remark: the discourse topic being the nature of the timing of your parents' return, a scale is constructed that characterizes the timing in terms of degree of favorability. This is a scale ranging from good (very favorable) to bad (much unwanted). The literal message describes their timing as *perfect*, so one point is set on the scale close to the "Good" edge. In reality, as we remember, and as your boyfriend would readily admit, their

timing was quite *horrible*. A second point is then set at the "Bad" edge (see Figure 2). Once both points are set, the distance between them is computed. It is a significant distance, which licenses the ironic meaning.

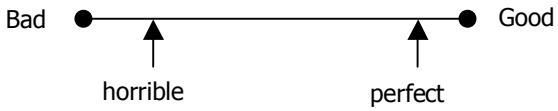


Figure 2: "What perfect timing!"

The present approach stems from the Indirect Negation theory of irony [3]. As such, it shares with classic pragmatic theory the classification of irony as a breach of a norm, but differs from it in its claim that the literal meaning is not discarded or rejected since it is crucial for computing the ironic meaning. Generalizing the Indirect Negation theory of irony, it is not required that the state of affairs designated by the ironic expression be an expected or desirable state of affairs. It makes no difference if the tone of the literal meaning is positive (as in ironic criticism) or negative (as in ironic praise), as long as it is located far enough from the implied meaning. Empirical experiments performed by Dukas [4] on visual irony in still and moving pictures corroborate our approach. Dukas has shown that the contrast between foreground (literal message) and background (implicated message) is more important in creating the irony than directionality, i.e. tone, of the two messages. Contrary to the predictions of the Indirect Negation theory of irony, he found that ironies in which the foreground was positive and the background was negative were not significantly easier to detect than ironies in which the messages were presented the other way around.

Scales provide additional insight in the account of ironic understatements and overstatements: an expression serving as an understatement in one context may function as an overstatement in an "opposite context", one in which the contextual-point lays at the opposite edge of the scale.

The scale model accounts elegantly for the occurrence of the so-called "ironic cues" typical of ironic statements. Hyperboles, intensifying adverbs, and intonation all serve to widen the gap between the literal and contextual points on the scale. By driving the literal point closer to an edge the distance between the two points is increased, giving rise to better irony.

A computational model for irony using LSA

From a computational point of view, the quantitative nature of the scale model suggests it can serve as a theoretical basis for a computational model of irony. The reasons for choosing LSA as the formal framework for our solution are a threefold: first, its latent contextual knowledge can be queried in order to extract relevant bits of information, namely the implicature. Furthermore, LSA provides a metric that can be utilized to calculate the distance between the implicature and the given literal utterance. Together, the contrast that lies in the heart of the irony can be computed. Lastly, the successful application of LSA as a model for metaphor [5] suggests it may play a role in a model for other types of non-literal language, and specifically for irony.

Latent Semantic Analysis, henceforth LSA, is a general theory of acquired similarity and knowledge representation. It is a bag-of-words model that ignores whatever linguistic structure is present in the text (morphological, syntactic, narrative, etc.) and is sensitive only to occurrences of words. The basic assumption of LSA is that words that have similar meanings tend to co-occur in texts. LSA's power lies in the fact that it is sensitive not only to direct co-occurrences, but can also infer indirect relations between words across texts. Similarity is measured in LSA as distance between vectors representing text items (or novel combinations of text items), defined as the cosine of the angle between them: the higher the cosine, the more similar the items.

In this work we attempt to utilize LSA for the task of automatic detection of irony. At first glance it seems to be the ideal model for irony: it has a metric for comparing sentences to sentences and words to words, it holds a representation of semantic relations between words, and it has shown "proof of concept" in many tasks that involve measures of similarity. However, LSA has its drawbacks. Two characteristic of LSA were taken into consideration before applying it to the task at hand. First, it is unable to distinguish synonyms from antonyms. Typical examples of irony that make use of antonyms ("Very funny", "What wonderful weather!") will go unnoticed. Second, it is ignorant of function words such as negation markers and intensifying adverbs, which are crucial clues in detecting irony. In light of these limitations, we decided to focus on a special type of irony, namely ironic headlines ("Afghanistan: a touristy leisure getaway"). These ironies can be expressed without negation markers and intensifying adverbs, and are typically based on the inappropriateness of concepts, not of antonyms.

Consider the following ironic headline:

Priorities[1]

"The most important thing in the world is eyebrow design"

Beauty queen and model Ilanit Levy (Yedioth Ahronoth)

Irony arises from the contrast between the meaning of the headline *Priorities* and the topic *eyebrow design*. Now suppose we replace *eyebrow design* by *buying a house* or *health*: the result is a literal and somewhat dull headline. We expect LSA to be sensitive to these differences.

The key idea in using LSA to detect irony is to look for dissimilarity and contrast, which in LSA means low similarity scores. Given a headline (*Priorities*) and a set of alternative topics (*eyebrow design*, *buying a house*, *health*), the model attempts to find the most ironic one by:

1. Computing the LSA similarity score between the headline and each of the alternative topics.
2. Ordering the alternatives according to their scores.
3. Outputting the pair headline-topic that received the lowest score as ironic.

We do not attempt in this work to cope with the more general detection task of judging for any arbitrary input utterance if it is ironic or not based on its content and the context, although it is definitely an interesting problem that should be addressed in the future.

In order to assess LSA's applicability to the task of irony detection based on the model of scales, we performed a series of tests. The main question we set out to answer was whether the proposed computational model mimics humans' behavior on tasks of irony detection.

Method

Materials and Procedure

Two irony detection tasks were presented to human subjects and to the computational model: two multiple choice tasks and a ranking task.

The multiple choice tasks consisted of 20 questions. Each question was presented as a set of alternative utterances, from which subjects were instructed to choose the most ironic one. In the main multiple choice task, 10 questions were presented with 4 alternatives each (Location items) - a total of 40 items. A second multiple choice task consisted of 10 questions with 2

[1] Appeared in the "Overheard" section of the Haaretz Magazine English edition, 28 February, 2003.

alternatives each (Government items) – a total of 20 items. An example of a Location question of this type with 4 alternatives (underline in the original):

a. Iceland is really polluted.

b. New York is really polluted.

c. Goa is really polluted.

d. Afghanistan is really polluted.

In the corresponding questions for the computational model, the alternatives were presented as pairs of the underlined elements in the questions presented to humans:

a. (Iceland, polluted)

b. (New York, polluted)

c. (Goa, polluted)

d. (Afghanistan, polluted)

The ranking task consisted of 7 questions, all based on real examples of ironic headlines from an Israeli newspaper[2]. The text under the headline included a blank, for which 3 alternative completions were given. Subjects were asked to rank the degree of irony for each alternative with respect to the headline on a scale of 1 (not ironic) to 10 (very ironic).

An example of a ranking question of this sort:

Priorities

"The most important thing in the world is_____"

[ ] health

[ ] eyebrow design

[ ] buying a house

In the corresponding questions for the computational model, the alternatives were presented as pairs of the headline and each of the alternative completions:

a. (priorities, health)

b. (priorities, eyebrow design)

c. (priorities, buying a house)

Participants

22 human subjects participated in the experiment: 52% female, 48% male, average age of 28.9 years. All were university students or graduates who volunteered to participate in the experiment.

[2] Examples were based on excerpts from sections of the Israeli Haaretz Magazine: "Kikar Ha-Medina" in the Hebrew edition (27 September, 2002; 28 February, 2003; 14 March 2003), and "Overheard" in the English edition (28 February, 2003). See Appendix A for the actual test items included in the analysis.

Simulations

LSA simulations were performed using the online web-based LSA application One-To-Many Comparison[3], on the General Reading up to 1st year college semantic space with 300 dimensions.

The questionnaire for human subjects was in Hebrew, and was translated to English for the evaluation of the computational model. It was verified that all words used in the questions existed in the corpus: if the word that appeared in the questionnaire for humans was not part of LSA's inventory, a near exact translation was used instead.

Results

We now present the results of the computer simulations in comparison to humans' responses. Results for the multiple choice task are presented separately for the Location items (Table 1) and for the Government items (Table 2). For each alternative, LSA similarity scores are shown above the percentage of participants who chose it as most ironic. Shaded in dark gray are humans' and the model's first choices of the most ironic alternative for each question. Shaded in light gray are the model's second choices that match humans' first choices.

	New York	Goa	Iceland	Afghanistan
desert	0.07 59.09%	-0.06 4.55%	-0.06 36.36%	0.37 0%
tourism	0.13 0%	0.06 4.55%	0.19 0%	0.03 95.45%
highrise	0.13 4.76%	-0.03 33.33%	0.1 9.52%	-0.01 52.38%
polluted	0.06 0%	0.05 4.55%	-0.01 95.45%	-0.01 0%
island	0.24 18.18%	0.07 9.09%	0.57 0%	0.08 72.73%
desolate village	0.07 95.45%	0.15 4.55%	0.06 0%	0.08 0%
romantic atmosphere	0.08 0%	0.01 0%	0.03 0%	0 100%
over populated	0.32 0%	0.12 13.64%	0.16 77.27%	0.23 9.09%
bustling metropolis	0.26 0%	-0.03 31.82%	0.03 45.45%	0.03 22.73%
modern	0.18 0%	0.02 4.55%	0.05 4.55%	0.1 90.91%

Table 1: Results for Location items

[3] Available at http://lsa.colorado.edu.

As can be seen in Table 1 above, when comparing the model's first choice with humans' first choice (shaded dark gray), the model got only 3 items, 30%, correct (7% corrected for guessing by the formula [correct-chance/1-chance][4]). However, on a more lax comparison taking into consideration the model's first and second choices (shaded light gray), the model got 9 items, 90%, correct (80% corrected for guessing).

	democracy	dictatorship
freedom of opinion	0.56 27.27%	0.31 72.73%
censorship	0.27 80.95%	0.26 19.05%
decentralization	0.43 4.55%	0.36 95.45%
secret police	0.13 100%	0.2 0%
human rights	0.3 4.55%	0.16 95.45%
centralization	0.41 95.45%	0.28 4.55%
political parties	0.66 9.52%	0.45 90.48%
free elections	0.51 9.09%	0.35 90.91%
rule of the people	0.43 13.64%	0.42 86.36%
violence	0.35 94.74%	0.33 5.26%

Table 2: Results for Government items

Comparing the model's first choice with humans' first choice (shaded dark gray) for Government items, the model got 7 items, 70%, correct (40% corrected for guessing).

Item analysis was performed to check the correlation between humans' judgment of irony and the model's judgment. Each pair of headline and topic received two scores: the number of participants that chose it as most ironic (out of 22), and a score of irony according to the model from 1 to 4, where 1 is least ironic and 4 is most ironic. Spearman correlation revealed a significant correlation between the two variables for the Location items: Spearman coefficient = 0.61, p=0.0001. In the Government items correlation was not found. In both

[4] Akin to the correction in Landauer & Dumais (1997) in evaluating LSA's success rate on the TOEFL synonymy test.

Location and Government items together a correlation of Spearman coefficient = 0.34, was found ($p<0.01$).

We turn now to the results of the ranking task. Recall that questions in this part were based on real examples of ironic headlines. Of the 7 questions in this part of the experiment, participants failed to detect the irony in one question, and it was not included in the comparison. An average of participants' rankings was calculated for each of the remaining six headline-completion pairs (10 - very ironic, 1 - not ironic). These averages are shown in Table 3, along with the LSA similarity scores for each pair.

priorities	eyebrow design	buying a house	health
	9.59 0.13 (**0.02**)	4.38 **0.08**	1.19 **0.15**
matchmaker	rapist	cashier	teacher
	9.36 -0.05	5.19 0.01	3.19 0.06
Judaism	death	height	success
	8.68 0.17 (**0.04**)	4.38 -0.01 (**0.06**)	1.19 0.01 (**0.31**)
Shakespeare	soap opera	story	drama
	7.73 0.28	3.62 0.13	3.19 0.84
nirvana	full volume	loud	quiet
	6.68 -0.01	5.19 0	1.81 0.09
profession	son	driver	consultant
	8.91 0.13 (**0.08**)	2.95 0.03 (**0.08**)	1.81 0.34 (**0.10**)

Table 3: Results for ranking task

The model did not succeed in mimicing humans' rankings: only in 2 out of the 6 questions (*matchmaker*, *nirvana*: marked with gradual shading) did the model mimic the scale given by humans for the alternative topics. In three cases (*priorities*, *profession*, *Shakespeare*) it did not succeed in detecting the most ironic alternative (*eyebrow design*, *son*, *soap opera* respectively). In the remaining question (*Judaism*) the model ranked the alternatives totally opposite to the participants. However, carefully varying the items presented to LSA had a drastic effect on the results, as indicated by the figures in boldface. These effects are described and discussed in the next section.

Discussion and conclusions

The results presented above provide supporting evidence for the model of scales, showing that humans' judgments of irony correlate with distances between concepts. However, they do not univocally support the viability of the proposed model of irony based on LSA. On the one hand, a significant substantial correlation was found between judgments of the model and humans in the main multiple choice task. On the other hand, in order to achieve a success rate of 80% on this task, we had to take into consideration both the model's first and second answers. Results in the second multiple choice task and in the ranking task were less encouraging.

However, we believe the model should be tested more thoroughly before a conclusion regarding its viability is reached. Firstly, varying the corpus on which LSA is trained may have a considerable effect on the results. For example, the real-life examples in the ranking task were taken from a contemporary Israeli newspaper. In order to fully appreciate these ironic headlines one must be knowledgeable about current political and social issues in today's Israel. The human participants were clearly knowledgeable in this respect, but the corpus LSA was trained on was not.

Secondly, we noticed that subtle changes in the input to LSA have drastic effects on the results (see figures in boldface in Table 3). Thus, using *job* instead of *profession* and *eyebrow* instead of *eyebrow design* in the ranking task resulted in different, and correct, rankings by the model (cosine(*job*, *son*)=0.08 cosine(*job*, *driver*)=0.08, cosine(*job*, *consultant*)=0.10, cosine(*priorities*, *eyebrow*)=0.02). Using *Atonement* instead of *Judaism* also brought out a correct scale: cosine(*Atonement*, *death*)=0.04, cosine(*Atonement*, *height*)=0.06, cosine(*Atonement*, *success*)=0.31. This variability in similarity scores demonstrates that LSA's judgments do not always match our intuitions: while we would judge *job* and *profession* as near synonyms, they behave differently in the semantic space; while we know irony results from *eyebrow* and not from *design*, LSA should be told this explicitly.

In conclusion, based on our findings we believe LSA can serve as a basis for a working model of irony. However, its limitations should be understood, and it should be augmented by mechanisms that are sensitive to negation markers, to intensifiers, and to the distinction between synonyms and antonyms. Further research should also explore the effects of changing the training corpus and methods for detecting irony without relying on a set of pre-defined alternatives.

References

[1] Landauer, Thomas K. and Susan T. Dumais (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review* 104(2): 211-240.

[2] Rubinstein, Aynat (2002). "Irony on a scale - A discourse theoretic account of irony", a talk given at GIM2002, German Israeli Minerva Summer School on Computational Linguistics.

[3] Giora, Rachel (1995). On Irony and Negation. *Discourse Processes* 19: 239-264.

[4] Dukas, Gideon (1997). On Aptness of Visual Irony: Testing Irony in Photography and Cinema. MA Thesis, Tel Aviv University.

[5] Kintsch, Walter (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review* 7: 257-266.

Appendix A

Following are the test items used in the ranking task. Originally in Hebrew, they are all based on excerpts from the Israeli Haaretz Magazine (see footnote 2 for details):

Priorities

"The most important thing in the world is_____"

[ ] health

[ ] eyebrow design

[ ] buying a house

Matchmaker

"I would have introduced her to John, who is a renown _____"

[ ] rapist

[ ] cashier

[ ] teacher

Judaism

"I wished that their kids _____, and I also prayed for it in the synagogue on The Day of Atonement"

[ ] would die

[ ] would be tall

[ ] would succeed in life

Shakespeare

"It's just a _____. After all, what is Romeo and Juliet?"

[ ] drama

[ ] story

[ ] soap opera

nirvana

"While he cooks he turns on the TV _____. He claims it calms him"

[ ] loud

[ ] quiet

[ ] full volume

Profession

"The prime minister's _____"

[ ] consultant

[ ] son

[ ] driver

A Cross-linguistic Study on Bilingual Terminology Acquisition from Comparable Corpora applicable to Figurative languages

Fatiha SADAT‡, Masatoshi YOSHIKAWA*, and Shunsuke UEMURA‡

‡ Graduate School of Information Science,
Nara Institute of Science and Technology (NAIST)
8916-5 Takayama, Ikoma. Nara 630-0101. Japan
* Information Technology Center, Nagoya University

Mail: fatia-s@is.aist-nara.ac.jp

Abstract

The present paper describes an approach to bilingual lexicon extraction from comparable news articles and evaluations on Cross-Language Information Retrieval. Our goal is to learn translation lexicons using scarce resources, i.e. resources available on the Internet. News articles are exploited for bilingual terminology acquisition. A combined translation model involving the corpora-based model, readily available bilingual dictionaries and transliteration of the special phonetic alphabet of foreign words and loanwords (here Japanese katakana), is proposed. Evaluations using large-scale test collection on Japanese-English and SMART retrieval system revealed the proposed combination of comparable corpora, bilingual dictionaries, and transliteration to be highly effective in Cross-Language Information Retrieval.

1 Introduction

Large text corpora represent a crucial resource for the acquisition of bilingual terminology and the enrichment of multilingual lexical resources. According to previous researches (Dagan, 1994; Dejean et al., 2002; Diab and Finch, 2000; Fung, 2000; Koehn and Knight, 2002; Peters and Picchi, 1995; Rapp, 1999; Shahzad and al., 1999; Tanaka and Iwasaki, 1996), the extraction of bilingual terminology showed a great success, especially when combining different models, involving bilingual dictionaries, corpora and possibly thesauri.

In the present paper, our goal is to learn translation lexicons using scarce resources, i.e. resources available on the Internet. We are concerned by exploiting news articles as comparable corpora in order to translate terms in a source language to any specified target language. The extracted source terms could be generalized or specialized including figurative forms such as metaphors, metonyms, idioms or ironic manipulations of their canonical forms and compounds. Our preliminary study was conducted on (Japanese, English) language pair using general-domain comparable corpora and could be extended to figurative languages. Evaluations were conducted on Cross-Language Information Retrieval (CLIR) using large-scale test collection for Japanese and English.

The remainder of the present paper is organized as follows: Section 2 presents an overview of the corpora-based approach for bilingual terminology acquisition. Section 3 describes the linear combination of different translation models involving comparable corpora, bilingual dictionaries

and transliteration. Experiments and evaluations in CLIR are presented in Sections 4. Section 5 introduces an application to figurative languages. Section 6 concludes the present paper.

2 An overview of the Proposed Approach

Unlike parallel texts, which are clearly defined as translated texts, there is a wide variation of non-parallel-ness in monolingual data. It can be manifested in the topic, the domain, the authors, the time period, etc. Comparable corpora are collections of texts from pairs or multiples of languages, which can be contrasted because of their common features. We rely on such comparable corpora for the extraction of bilingual terminology, in the form of translations and/or similar terms.

We follow the model proposed by (Fung, 2000; Rapp, 1999; Dejean et al., 2002). First, word frequencies, context word frequencies in surrounding positions (here three-words window) are computed following statistics-based metrics. Context vectors for each term in the source language and the target language are constructed. We use the *log-likelihood ratio* (Dunning, 1993) as expressed in equation (1).

$$2\log \lambda = K_{11}\log \frac{K_{11}N}{C_1R_1} + K_{12}\log \frac{K_{12}N}{C_1R_2} + K_{21}\log \frac{K_{21}N}{C_2R_1} + K_{22}\log \frac{K_{22}N}{C_2R_2} \qquad (1)$$

Where, $C_1 = K_{11} + K_{12}$, $C_2 = K_{21} + K_{22}$,
$R_1 = K_{11} + K_{21}$, $R_2 = K_{12} + K_{22}$,
$N = K_{11} + K_{12} + K_{21} + K_{22}$,
K_{11} = *frequency of common occurrences of word* w_i *and word* w_j,
K_{12} = *corpus frequency of word* w_i, $- K_{11}$,
K_{21} = *corpus frequency of word* w_j $- K_{11}$,
$K_{22} = N - K_{12} - K_{22}$.

Next, context vectors of the target words are translated using a preliminary seed lexicon. We consider all translation candidates, keeping the same context frequency value as the source term. This step requires a seed lexicon that will be enriched using the proposed bootstrapping approach of this paper.

Similarity vectors are constructed for each pair of source term and target term using the *cosine metrics* (Salton, 1983), as expressed in equation (2).

$$Similarity\ (term_s, term_t) = \frac{\sum_k v_{sk} v_{tk}}{\sqrt{\sum_k v_{sk}^2 \sum_k v_{tk}^2}} \qquad (2)$$

Where, v_{ik} represents co-occurrence frequencies in context vectors of the source term $term_s$ with term $term_k$. and v_{jk} represents co-occurrence frequencies in context vectors of the target term $term_t$ with term $term_k$.

Thus, similarity vectors are constructed to yield a probabilistic translation model $P_{comp}(t|s)$.

3 Linear Combination of Different Translation Models

Combining different models has showed success in previous research (Dejean et al., 2002). We propose a combined model involving comparable corpora, readily available bilingual dictionaries as well as transliteration for the special phonetic or spelling representation of Japanese language (represented by Katakana alphabet).

3.1 Dictionary-based Translation Model

General-purpose dictionaries are basic source for translations and could be exploited for bilingual terminology extraction. The proposed dictionary-based translation model is derived directly from readily available bilingual dictionaries, by considering for each source entry all translation candidates and their associated phrases.

If a source terms appears with *N* translation alternatives in the bilingual dictionary; thus, for each pair of source term *s* and its target translation *t*, the -probability $P_{dict}(t|s)$ is computed as follows:

$$P_{dict}(t \mid s) = \begin{cases} \frac{1}{N} & if \quad N > 0 \\ 0 & else \end{cases} \qquad (3)$$

3.2 Transliteration Model

Transliteration is the phonetic or spelling representation of one language using the alphabet of another language. The special phonetic alphabet (here Japanese katakana) to foreign words and loanwords requires *romanization* or transliteration (Knight and Graehl, 1998). Japanese vocabulary is frequently imported from other languages, primarily (but not exclusively) from English. *Katakana*, the special phonetic alphabet is used to write down foreign words and loanwords, example names of persons and other terms. The English word *computer* is transliterated in Japanese katakana as コンピューター, as well *engineer* is transliterated as エンジニアー, and *space shuttle* is transliterated as スペースシャトル. Named entities such as proper names of foreign (else than Japanese) persons, locations and organizations, are transliterated in Japanese. An example is *Bill Clinton* as named entities, which is transliterated in Japanese as ビルクリントン.

Assume a source term *s* (written in katakana) is represented by *N* transliteration alternatives. Each transliteration *t* will be represented by a probability $P_{translit}(t|s)$ as follows:

$$P_{translit}(t \mid s) = \frac{1}{N} \qquad (4)$$

In the present paper, *KAKASI*[1]a language processing inverter available on the Internet is used to convert terms written in katakana to their romaji forms, i.e., the alphabetical description of Japanese pronunciation and thus complete a transliteration. Note that most Japanese terms presented to KAKASI system in katakana showed a unique transliteration; in this case $P_{translit}(t|s)$ is equal to 1.

[1] http://kakasi.namazu.org

3.3 Linear Combination

The combined probabilistic lexical model is represented by three translation sub-models: the comparable corpora-based sub-model represented by $P_{comp}(t|s)$, the bilingual dictionary-based sub-model represented by $P_{dict}(t|s)$ and the transliteration sub-model represented by $P_{translit}(t|s)$.

Translation alternatives are ranked according to the combined probability. A fixed number of top-ranked translation candidates are selected and misleading candidates are discarded.

4 Experiments and Evaluations

Experiments have been carried out to measure the improvement of our proposal on bilingual Japanese-English tasks in CLIR, i.e. Japanese queries to retrieve English documents.

4.1 Linguistic Resources

- A collection of news articles from *Mainichi Newspapers* (1998-1999) for Japanese and *Mainichi Daily News* (1998-199) for English was considered as comparable corpora, because of their common feature of the time period.
- Morphological analyzers, *ChaSen*[2] version 2.2.9 for texts in Japanese and *OAK*[3] for English texts were used in linguistic pre-processing.
- *EDR* (EDR, 1996) and *EDICT*[4] bilingual Japanese-English dictionaries were used for translation.
- *KAKASI*[5], a language processing inverter and free software, available on the Internet was used in the transliteration model.
- *NTCIR*[6], a large-scale test collection was used to evaluate the proposed strategies in CLIR.
- *SMART*[7] information retrieval system (Salton, 1971), which is based on vector model, was used to retrieve English documents.

4.2 Evaluation and Results

Content words (nouns, verbs, adjectives, adverbs) were extracted from English and Japanese corpora. In addition, foreign words (mostly represented in katakana) were extracted from Japanese texts. Thus, context vectors were constructed for 13,552,481 Japanese terms and 1,517,281 English terms. Similarity vectors were constructed for 96,895,255 (Japanese, English) pairs of terms.

We conducted experiments and evaluations on the monolingual and bilingual tasks of NTCIR test collection. Topics 0101 to 0149 were considered and key terms contained in the fields, title

[2] http://chasen.aist-nara.ac.jp/
[3] http://nlp.cs.nyu.edu/oak/
[4] http://www.csse.monash.edu.au/~jwb/wwwjdic.html
[5] http://kakasi.namazu.org/
[6] http://research.nii.ac.jp/ntcir/
[7] ftp://ftp.cs.cornell.edu/pub/smart

<*TITLE*>, description <*DESCRIPTION*> and concept <*CONCEPT*> were used to generate 49 queries in Japanese and English.

Results and performances of different translation models and their combination are described in Table 1. The combined dictionary-based and transliteration model *'DT'* showed 84.94% improvement of the monolingual retrieval, while the comparable corpora-based model '*SCC*' showed a lower improvement in average precision compared to the monolingual retrieval and the combined dictionary-based and transliteration model *'DT'* with 52.81% of the monolingual retrieval. The proposed combination of comparable corpora, bilingual dictionaries and transliteration '*DT&SCC'* showed the best performance in terms of average precision with 88.18% of the monolingual counterpart, +3.82% compared to the dictionary-based method and +66.97 compared to the comparable corpora model taken alone.

Table 1. Results and evaluations on different translation models and their combination using NTCIR test collection

Method			Average Precision	% Monolingual	% Difference Improvement		
ME	-	Monolingual English	**0.2683**	**100**	--		
DT	-	Dictionary and Transliteration	0.2279	84.94	- 15.05	--	
SCC	-	Comparable Corpora	0.1417	52.81	- 47.18	-37.82	--
DT&SCC	-	Combination	**0.2366**	**88.18**	**-11.81**	**+3.82**	**+66.97**

5 Application to Figurative languages

A figurative language has considerable expressive power that is matched by a potential for misunderstanding and so it must be used judiciously. Knowledge acquisition from comparable corpora of figurative languages could contribute as entries in a bilingual dictionary, although this task is quite complex because of figurative senses and the context-dependence that a word could be related to.

Bilingual terminology acquisition from comparable corpora might be faced with cultural incompatibilities that emerge on the level of figurative language. Figurative forms such as metaphors, metonyms, idioms or ironic manipulations of their canonical forms and compounds as well as dialect, slang words and technical terms outside their normal scope are too risky; unless these forms are widespread among languages. The problem of *Cultural differences* should be studied more deeply and solutions for their translations across languages should be found.

At this stage, we did not evaluate the proposed strategy of bilingual terminology acquisition from comparable corpora on figurative languages, but we think that is very interesting to include in future research. Thesauri or ontologies could help to determine the context of a word and thus contribute in the combined translation model for figurative languages.

6 Conclusion

We investigated the approach of extracting bilingual terminology from comparable corpora-based for (Japanese, English) language pair. A combined model involving comparable corpora, readily available bilingual dictionaries and transliteration was found very efficient and could be used to enrich bilingual lexicons and thesauri. Most of the selected terms were considered as translation candidates or expansion terms in CLIR. Exploiting different translation models revealed to be highly effective. Ongoing research is focused on studies and applications on figurative languages, solutions to Word Sense Disambiguation and more strategies to fulfill needs of Cross-Language Information Retrieval.

References

Dagan, I., Itai, I. : Word Sense Disambiguation using a Second Language Monolingual Corpus. Computational Linguistics, Vol. 20-4 (1994) 563-596.

Dejean, H., Gaussier, E., Sadat, F.: An Approach based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction. In Proceedings of the 19th International Conference on Computational Linguistics COLING (2002) 218-224.

Diab, M., Finch, S.: A Statistical Word-Level Translation Model for Comparable Corpora. In Proceedings of the Conference on Content-based Multimedia Information Access RIAO (2000).

Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence. Computational linguistics, Vol. 19-1 (1993) 61-74.

EDR: Japan Electronic Dictionary Research Institute, Ltd. EDR electronic dictionary version 1.5 technical guide. Technical report TR2-007, Japan Electronic Dictionary research Institute, Ltd (1996).

Fung, P.: A Statistical View of Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. In Jean Véronis, Ed. Parallel Text Processing (2000).

Knight, K., Graehl, J.: Machine Transliteration. Computational Linguistics, Vol. 24-4 (1998).

Koehn, P., Knight, K.: Learning a Translation Lexicon from Monolingual Corpora. In Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition (2002).

Peters, C., Picchi, E.: Capturing the Comparable: A System for Querying Comparable Text Corpora. In Proceedings of the 3rd International Conference on Statistical Analysis of Textual Data (1995) 255-262.

Rapp, R.: Automatic Identification of Word Translations from Unrelated English and German Corpora. In proceedings of the European Association for Computational Linguistics (1999).

Salton, G.: Automatic Processing of Foreign Language Documents. Journal of the American Society of Information Science (1970) 187-194.

Salton, G., McGill, J.: Introduction to Modern Information Retrieval. New York, Mc Graw-Hill (1983).

Shahzad, I., Ohtake, K., Masuyama, S., Yamamoto, K.: Identifying Translations of Compound Using Non-aligned Corpora. In Proceedings of the Workshop MAL (1999) 108-113.

Tanaka, K., Iwasaki, H.: Extraction of Lexical Translations from Non-Aligned Corpora. In proceedings of the 13th International Conference on Computational Linguistics COLING (1996).