Corpus-driven Metaphor Harvesting

Astrid Reining
Institute of Romance Languages
University of Hamburg
20146 Hamburg, Germany
astrid.reining@uni-hamburg.de

Birte Lönneker-Rodman
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA
loenneke@icsi.berkeley.edu

Abstract

The paper presents a corpus-based method for finding metaphorically used lexemes and prevailing semantico-conceptual source domains, given a target domain corpus. It is exemplified by a case study on the target domain of European politics, based on a French 800,000 token corpus.

1 Introduction

This investigation is situated within the framework of the Hamburg Metaphor Database[1] (HMD) (Lönneker and Eilts, 2004), which collects manual annotations of metaphors in context. HMD annotation terminology refers to cognitive linguistic accounts of metaphor. These suggest that abstract "target" concepts are often thought and talked of in terms of less abstract "source" concepts (Section 2). On these accounts, the paper presents a method for finding metaphorically used lexical items and characterizing the conceptual source domains they belong to, given a target domain corpus.

After mentioning related work on metaphor annotation (Section 3), we exemplify our method by a case study on the target domain of European politics, for which a French 800,000 token corpus is prepared and imported into a corpus manager (Section 4). Using corpus manager functions, a small set of highly salient collocates of *Europe* are classified as candidates of metaphorical usages; after assessing their metaphoricity in context, these lexemes are grouped into semantico-conceptual domains for which, in a final step, additional lexical instantiations are searched (Section 5). Two important source domains (BUILDING and MOTION) are detected, which are supported by over 1,000 manual corpus annotations. The domains can be characterized as small networks of EuroWordNet synsets (nodes) and lexical as well as conceptual relations (Section 6). Section 7 concludes the paper.

2 Theoretical Aspects

The Conceptual Theory of Metaphor (CTM) worked out originally by (Lakoff and Johnson, 1980) claims that conceptual metaphors such as GOOD IS UP and TIME IS MONEY structure the way we think and influence the way we use language. Conceptual metaphors are mappings between conceptual domains, for example between the target domain GOOD and the less abstract source domain UP, or between TIME (target) and MONEY (source).

Conceptual metaphors are rarely *directly* referred to in speech or writing: Whereas *time is money* is a standing expression in English, this is much less so for many other conceptual mappings (cf. *?good is up*). Consequently, corpus analysis cannot have as a goal finding conceptual mappings as such. Rather, it can find their manifestations through non-literal usages of lexical items – i.e., contexts in which source domain words are used to refer to elements in the target domain.

For example, *high* (a word from the UP source domain) means 'good' in the expression *high marks*; and *spend* or *save*, used in the source domain to refer to actions involving money, refer to actions in the

[1] http://www1.uni-hamburg.de/metaphern

Proceedings of the Workshop on Computational Approaches to Figurative Language, pages 5–12,
Rochester, NY, April 26, 2007. ©2007 Association for Computational Linguistics

target domain of TIME when used in contexts such as *spend time* or *save time*.

Adopting a broad notion of metaphor based on CTM, we refer to such non-literal usages (though often conventionalized) as *lexical metaphors* in this paper. Prominent conceptual metaphors are illustrated by a larger number of lexical metaphors, which support the systematicity of their mapping.

3 Related Work

Earlier projects annotating metaphor in corpora include (Martin, 1994) and (Barnden et al., 2002). In what follows, we give two examples of recent work.

Gedigian et al. (2006) annotated a subset of the *Wall Street Journal* for the senses of verbs from Motion-related, Placing, and Cure frames which were extracted from FrameNet (Fillmore et al., 2003). The annotation shows that more than 90% of the 4,186 occurrences of these verbs in the corpus data are lexical metaphors in the above sense. Gedigian et al. (2006) conclude that in the domain of economics, Motion-related metaphors are used conventionally to describe market fluctuations and policy decisions. A classifier trained on the annotated corpus can discriminate between literal and metaphorical usages of the verbs.

Lee (2006) compiled a 42,000 word corpus of transcribed doctor-patient dialogues, exhaustively hand-annotated for stretches of metaphorical language. These are provided with conceptual labels enabling the author to identify prevalent and interrelated metaphorical mappings used as part of communicative strategies in this domain.

4 The European Constitution Corpus

Exploration and annotation of a corpus to find information regarding its predominant conceptual source domains is most productive when applied to an abstract and novel target domain. Abstractness calls for ways to make the topic cognitively accessible, and novelty entails a certain openness about the particular source domains that might be activated for this purpose.

Abstractness and novelty are criteria fulfilled by the target domain selected for our study: European Constitutional politics. The domain is represented by the public discourse on the possible introduction of a European Constitution and on the corresponding French referendum (29 May 2005). The referendum allowed voters to accept or refuse the proposed Constitution text (the result being refusal). The remainder of this section describes the sources of the corpus (4.1), its acquisition (4.2), and pre-processing (4.3).

4.1 Sources

The corpus consists of two sub-corpora, collected from online versions of two French dailies, *Le Monde* and *Le Figaro*. The site `lemonde.fr` contains each article published in the printed version of the socialist-liberal newspaper *Le Monde*, whereas `lefigaro.fr` contains articles from the conservative newspaper *Le Figaro*.

4.2 Collection

From 27 April to 5 June, 2005, the above mentioned web sites were screened for articles on Europe and the European Constitution on a daily basis. For the case study presented in this paper, only articles dealing with the Constitution and discussing the referendum are retained. Each of these articles is a document of the European Constitution corpus and contains information on its publication date, author, and newspaper section (e.g. editorial). The selection of relevant articles is performed manually. This is labor-intensive but keeps noise to a minimum. As a guideline for distinguishing between "general" European topics and the referendum on the European Constitution, key words including *(European) Constitution* and *referendum* are used.

4.3 Preprocessing

The collected documents are converted into text format and annotated with a simple SGML tagset representing document meta data (in the header), paragraph boundaries, and sentence boundaries. Sentence detection is performed reusing TreeTagger scripts[2] because we POS-tag and lemmatize the texts using the TreeTagger (Schmid, 1994) and its French parameter file (Stein and Schmid, 1995). Finally, the corpus is verticalized for use with the Manatee/Bonito corpus manager (Rychlý and Smrž,

[2] Tokenizer perl script for modern French, available on Achim Stein's web page, `http://www.uni-stuttgart.de/lingrom/stein/forschung/resource.html` [accessed 4 September 2006].

2004), run in single platform mode on a Linux computer.

Table 1 gives an overview of the two sub-corpora. When collecting the corpus, relevance to the topic had been our only criterion. Interestingly, the two newspaper corpora are very similar in size. This means that the selected topic was assigned equal importance by the different newspaper teams. Tables 2 and 3 show absolute frequencies of the top ten lemmas, filtered by a list of 725 French stop words[3] but still including *oui* - 'yes' and *non* - 'no', buzz-words during the political debate on the European Constitution. The frequent words also give an impression of the domain centeredness of the corpus.

	Le Monde	**Le Figaro**
Size (tokens)	411,066	396,791
Distinct word forms	23,112	23,516
Distinct lemmas	13,093	13,618
Documents	410	489
Paragraphs	7,055	6,175
Subdocuments	59	n.a.
Sentences	17,421	17,210

Table 1: Size of the European Constitution corpus.

5 Lexical Metaphors and Source Domains

Our aim is to determine empirically salient metaphorical source domains used in the target domain of European politics, combined with the practical interest in speeding up the detection and annotation of lexical metaphors. In Subsection 3 above, two approaches to corpus annotation for metaphor were mentioned. Due to the size of the corpus and limited annotator resources, we cannot follow the full-text annotation approach adopted by Lee (2006). Neither do we proceed as Gedigian et al. (2006), because that approach pre-selects source domains and lemmas. In our approach, we search for metaphorically used lexical items from initially unknown source domains, so interesting lemmas cannot be listed *a priori*.

Therefore, we developed a new method which makes efficient use of existing corpus manager functions. The only constant is the representation of the target domain, predefined at a high level by the selection of our corpus. We fixed the lemma *Europe* as a low-level anchor of the target domain.[4] The investigation proceeds in three steps:

	Lemma	**Occurrences**
1.	*européen* - 'European'	2,033
2.	*non* - 'no'	2,306
3.	*Europe* - 'Europe'	1,568
4.	*politique* - 'political; politics'	1,159
5.	*oui* - 'yes'	1,124
6.	*France* - 'France'	1,110
7.	*constitution* - 'Constitution'	1,099
8.	*traité* - 'treaty'	906
9.	*monsieur* - 'mister'	872
10.	*mai* - 'May'	781

Table 2: Frequent words in the *Monde* sub-corpus.

	Lemma	**Occurrences**
1.	*européen* - 'European'	2,148
2.	*non* - 'no'	1,690
3.	*Europe* - 'Europe'	1,646
4.	*France* - 'France'	1,150
5.	*politique* - 'political; politics'	969
6.	*constitution* - 'Constitution'	921
7.	*oui* - 'yes'	917
8.	*ministre* - 'minister'	885
9.	*traité* - 'treaty'	856
10.	*devoir* - 'have to; obligation'	817

Table 3: Frequent words in the *Figaro* sub-corpus.

1. Statistically weighted lists of collocates of the target domain lemma *Europe* are calculated and screened for candidates of metaphorical language use (5.1).

2. For the obtained candidate collocates, the corpus is concordanced in order to discriminate usages and assign a source domain to each collocate (5.2).

3. The source domains are extended lexically, making use of EuroWordNet synsets and relations (5.3).

Corpus data drives the discovery of relevant lemmas in step 1. In steps 2 and 3, the corpus is used to increasingly refine and evaluate findings regarding relevant lemmas and source domains.

5.1 Collocate analysis

At this stage, it is necessary to set a range (span) within which candidate lemmas are to appear, mea-

[3] Developed by Jean Véronis: `http://www.up.univ-mrs.fr/veronis/data/antidico.txt` [accessed 4 September 2006].

[4] We could have started with a larger set of target domain lemmas, e.g. *européen* - 'European', *Bruxelles* - 'Brussels', *UE* - 'EU' etc. However, the results for *Europe* quickly proved to be sufficient in number and variety to illustrate the method.

sured in lemma counts starting with the anchor word *Europe*. Sample concordances show that *Europe* is often preceded by an article and sometimes by an additional preposition. Based on this insight, we heuristically restrict the context range for collocates to four (i.e. three words are allowed to occur between it and *Europe*). For example, *mère* ‘mother’ in Example (1) is retained as a collocate:

(1) Parce qu’elle a été la **mère**$_4$ fondatrice$_3$ de$_2$ l$_1$’**Europe** unie. (‘Because she [i.e. France] has been the founding mother of the unified Europe.’)

The minimum absolute frequency of the collocate within the specified context range is set to 3, which ensures results of at least three example sentences per co-occurring lemma. Intentionally, no restriction is applied to the part of speech of the collocate.

For both sub-corpora, lists of the top 100 collocate lemmas for *Europe* are calculated in the Manatee/Bonito corpus manager. We use the MI-score for ranking; it is based on the relative frequency of the co-occurring lemmas. Choosing MI-score over T-score is driven by an interest in salient collocates of *Europe*, whether or not they are common in the entire corpus. (T-score would tend to prefer collocates that occur frequently throughout the corpus.) The top collocates and their MI-scores are given in Tables 4 and 5.

MI-scores of the 100 top-ranked collocates are between 7.297 and 4.575 in the *Monde* corpus and between 7.591 and 4.591 in the *Figaro* corpus. Empirically, a threshold of $MI >= 6$ retains the most salient collocates of *Europe* in both corpora. These are 26 collocate lemmas from *Le Monde* and 20 from *Le Figaro*.

	Lemma	**MI**	**Abs. f**
1.	*panne* - ‘breakdown’	7.297	6
2.	*uni* - ‘unified’	7.275	13
3.	*réveil* - ‘awakening; alarm clock’	7.034	3
4.	*unification* - ‘unification’	6.864	4
5.	*paradoxe* - ‘paradox’	6.812	3
6.	*construire* - ‘construct’	6.799	31
7.	*résolument* - ‘decidedly’	6.619	3
8.	*otage* - ‘hostage’	6.619	3
9.	*utopie* - ‘utopia’	6.619	3
10.	*défier* - ‘defy, challenge’	6.619	3
...	...	...	...
26.	*révolte* - ‘revolt’	6.034	3
...	...	...	...
100.	*maintenant* - ‘now’	4.575	6

Table 4: Collocates of *Europe* in *Le Monde*.

	Lemma	**MI**	**Abs. f**
1.	*oriental* - ‘oriental, east’	7.591	8
2.	*unifier* - ‘unify’	7.498	6
3.	*Forum* - ‘Forum’	7.176	3
4.	*occidental* - ‘occidental, west’	7.065	5
5.	*panne* - ‘breakdown’	6.913	8
6.	*ouest* - ‘west’	6.691	3
7.	*prospère* - ‘prosperous’	6.591	4
8.	*bouc* - ‘goat’	6.498	3
9.	*patrie* - ‘fatherland, home country’	6.498	3
10.	*ruine* - ‘ruin’	6.498	3
...	...	...	...
20.	*doter* - ‘endow’	6.006	8
...	...	...	...
100.	*attacher* - ‘attach’	4.591	3

Table 5: Collocates of *Europe* in *Le Figaro*.

These highly salient collocates are evaluated for the potential of being used metaphorically in the target domain. The guideline underlying this evaluation is as follows: Those lexemes which, in at least one of their usages, designate entities belonging to domains more concrete than POLITICS (for example, BUILDING or FAMILY) are likely to be used metaphorically in the corpus. Specifically, among those collocates with $MI >= 6$, we identify the following metaphor candidates:

Le Monde *panne* - ‘breakdown’, *réveil* - ‘awakening; alarm clock’, *construire* - ‘construct’, *otage* - ‘hostage’, *bâtir* - ‘build’, *mère* - ‘mother’, *révolte* - ‘revolt’;

Le Figaro *panne*, *bouc* - ‘goat’, *ruine* - ‘ruin’, *traverser* - ‘traverse’, *racine* - ‘root’, *visage* - ‘face’, *reconstruire* - ‘reconstruct’.

Merging the lists yields 13 distinct candidate words, which are now evaluated based on contexts from within the corpus. There are a total of 112 occurrences of these lemmas co-occurring with *Europe* in a range of 4, the setting used to calculate collocate lists. Each of them is inspected in a context of at least one sentence. An annotator decides whether the usage is metaphorical, and confirms this in almost all of the cases (cf. Table 6).

5.2 Source domain identification

While disambiguating the 13 candidate lemmas in context, the annotator also assigns a source domain

	Monde	Figaro	Total	Metaphor
construire	31	13	44	44
reconstruire	0	3	3	3
bâtir	5	1	6	6
ruine	0	3	3	0 or 3
panne	5	7	12	12
traverser	2	7	9	9
mère	3	1	4	4
racine	2	5	7	7
visage	2	5	7	7
réveil	3	0	3	3
révolte	3	0	3	3
otage	3	2	5	5
bouc	3	3	6	6
Total	62	50	112	109 or 112

Table 6: Co-occurrences of candidate lemmas.

label to each occurrence. Actually, to hold the status of source domain in a conceptual mapping, a conceptual domain should be instantiated systematically by a number of lexical metaphors. Therefore, as long as this systematicity has not been verified, the assigned source domains are tentative.

Four tentative source domains are postulated, two of which might need to be split into subdomains. The general domains are BUILDING, MOTION, FIGHT, and LIVING BEING. Verbs *(.V)* and nouns *(.N)* instantiating them are listed in Table 7. The table also contains further (though still ambiguous) lemmas from the Top-100 collocate list supporting the source domains. Observations regarding the source domains, based on the 112 annotated lexical metaphors, are summarized in what follows.

The BUILDING source domain has the highest number of lexical metaphor instantiations. The ambiguity of *ruine* - 'ruin', however, is unresolvable: The texts talk about "ruins of Europe" after World War II; if understood as "ruins of cities/buildings in Europe," all of these occurrences are literal, but if interpreted as "ruins of the European political system," all of them are metaphorical. The ambiguity might be deliberate.

	Domain	**Disambiguated Lemmas**	**Futher collocates** (Top 100)
1.	BUILDING	*construire.V*, *reconstruire.V*, *bâtir.V*, *ruine.N ?*	*maison.N* - 'house', *fonder.V* - 'found'
2.	MOTION		
	– FORWARD MOTION	*panne.N*, *traverser.V*	*progresser.V* - 'progress', *avancer.V* - 'advance'
	– MOTOR VEHICLE	*panne.N*	*moteur.N* - 'motor'
3.	FIGHT	*otage.N*, *révolte.N*	*lutter.V* - 'fight'
4.	LIVING BEING		
	– PROCREATION	*mère.N*, *racine.N*	*père.N* - 'father', *naître.V* - 'be born'
	– BODY	*visage.N*	*dos.N* - 'back', *coeur.N* - 'heart'
	– REST	*réveil.N*	–

Table 7: Tentative source domains.

Also the MOTION domain has been assigned to a large number of disambiguated occurrences. The noun *panne* - 'breakdown' might instantiate a subdomain, such as (MOTION IN A) MOTORIZED VEHICLE; in some cases, it has been assigned MACHINE as source domain, purposely underspecified as to its motion-relatedness.

The LIVING BEING source domain is multifaceted, comprising PROCREATION, BODY, and REST, obviously personifying Europe. However, the frequency of lexical metaphors in these domains is in large part due to recurring quotations: For example, *mère* - 'mother' is used exclusively within the expression *la mère fondatrice de l'Europe* - 'the founding mother of Europe,' attributed to J. L. Rodriguez Zapatero; and *réveil* - 'awakening; alarm clock' (pointing to an action of a living being) occurs only as part of the expression *sonner le réveil de l'Europe* - 'ring the awakening/alarm of Europe,' coined by Ph. de Villiers. Finally, *bouc* - 'goat' is always part of the idiom *le bouc émissaire* - 'scapegoat'. Although it could be grouped under LIVING BEING, this expression is based on particular cultural knowledge rather than on systematic exploitation of general world knowledge about the source domain.

The FIGHT domain has the lowest count of lexical metaphors in the annotated co-occurrences of *Europe*. Also, the noun *otage* - 'hostage' occurs three times out of five within the expression *(ne pas) prendre l'Europe en otage* - '(not) take Europe hostage,' coined by N. Sarkozy and quoted as such.

To summarize, we observe that the most salient lexical metaphors co-occurring with *Europe* in the European Constitution corpus either refer to the source domains of BUILDING or MOTION, well-known source domains of conventional metaphors, or the lexical metaphors are sparse, referring to much less clearly delimited source domains such as LIVING BEING or FIGHT. Within the second group, there are a number of newly coined expressions, “one shot rich image metaphors,” (Lakoff, 1987) which evoke entire scenes but do not necessarily contribute to a wide-spread systematic exploitation of the source domain.

5.3 Lexical extension

Corpus annotation is now extended to a larger list of lemmas from the source domains of BUILDING and MOTION. The challenge here is finding additional lemmas that might exploit the postulated mappings, given a small set of disambiguated lemmas and ambiguous collocates (cf. Table 7). A lexical resource for French containing information on conceptual domains would be helpful here. EuroWordNet (EWN) could go in this direction. It defines many relation types, including the synonym relation inside synsets, as well as hyponym, near-antonym and meronym relations between synsets. Apart from these lexical relations, EWN also recognizes a family of semantico-conceptual INVOLVED relations, which relate a verb synset Y to a noun synset X if “X is the one/that who/which is *typically* involved in Ying” (Vossen, 1999) (our emphasis). Unfortunately, there are almost no actual instantiations of INVOLVED relations in the French part of EWN.

Taking our previously identified collocates of *Europe* as seeds, we extend our lemma list resorting to EuroWordNet synsets, as follows:

- lemmas in synsets lexically related by EWN relations to synsets containing our seed lemmas (hypo-, hyper-, anto-, mero- and synonyms);

- lemmas in synsets lexically related across part of speech to synsets containing our seed lemmas, by adding missing XPOS_NEAR_SYNONYM and XPOS_NEAR_ANTONYM relations ourselves;

- lemmas in synsets that are conceptually related to the seed synsets, by adding INVOLVED relations ourselves.

A reiteration of these steps (using encountered lemmas as new seeds) could lead very soon to general or peripheral lemmas. Ideally, one would set up a limit of reiteration per operation and consider all encountered lemmas as possible keywords of the domain. However, annotator resources being limited, we reduced the list of key lemmas to about 20 per domain (22 for BUILDING and 19 for MOTION), using human judgment.

At this stage, the restriction on the keyword of being a collocate of *Europe* is lifted. This results in search, disambiguation, and annotation being performed on *the entire corpus*. The annotator finds 663 lexical metaphors among the 1,237 occurrences of 22 BUILDING keywords, and 409 lexical metaphors among the 1,307 occurrences of 19 MOTION keywords. Each key lemma contributes positively to the count of lexical metaphors. Two consequences follow from these figures:

1. Both postulated source domains are systematically exploited by lexical metaphors.

2. Every second or third investigated occurrence is a lexical metaphor.[5] Collection and annotation of metaphors can thus proceed considerably faster on the key lemmas than it would on full text or randomly selected sentences.

For each lexical metaphor, the annotator provides EuroWordNet synset information. For the actual meaning in context, the synset belonging to the target domain is encoded. Additionally, the synset containing the metaphorically used lexeme *in its source domain sense* is indicated (“source synset”).

6 Source domain structure

The information on source synsets underlies conceptual maps of the two source domains. This is exemplified here by Figure 1, which represents the MOTION domain. Lexical metaphors are prefixed by M_; those word senses not encoded in EWN are marked with an asterisk at the end. Synsets shaded gray in Figure 1 contain at least one lemma that is exploited as a lexical metaphor, and as such attested in the European Constitution corpus. Ovals represent verb synsets, boxes show noun synsets, and hexagons depict events.

Relations between synsets illustrate the internal structure of the domain. Solid lines represent relations encoded in EuroWordNet. For legibility reasons, labels of hyponym relations have been omitted.

[5] In the vicinity of *Europe*, the ratio continues to be higher, with at least three quarters of the contexts being metaphorical.

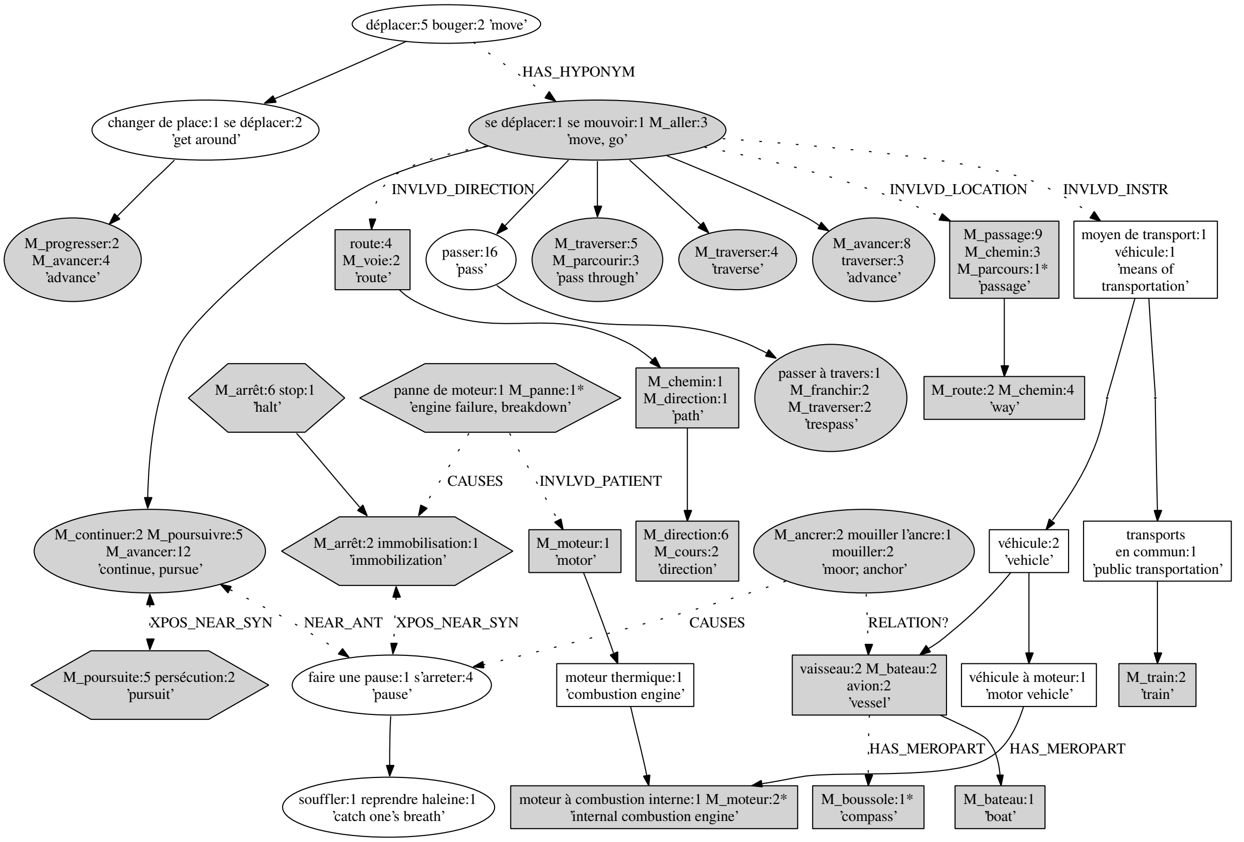
déplacer:5 bouger:2 'move'
HAS_HYPONYM
changer de place:1 se déplacer:2 'get around'
se déplacer:1 se mouvoir:1 M_aller:3 'move, go'
INVLVD_DIRECTION
INVLVD_LOCATION
INVLVD_INSTR
M_progresser:2 M_avancer:4 'advance'
route:4 M_voie:2 'route'
passer:16 'pass'
M_traverser:5 M_parcourir:3 'pass through'
M_traverser:4 'traverse'
M_avancer:8 traverser:3 'advance'
M_passage:9 M_chemin:3 M_parcours:1* 'passage'
moyen de transport:1 véhicule:1 'means of transportation'
M_arrêt:6 stop:1 'halt'
panne de moteur:1 M_panne:1* 'engine failure, breakdown'
M_chemin:1 M_direction:1 'path'
passer à travers:1 M_franchir:2 M_traverser:2 'trespass'
M_route:2 M_chemin:4 'way'
CAUSES
INVLVD_PATIENT
M_continuer:2 M_poursuivre:5 M_avancer:12 'continue, pursue'
M_arrêt:2 immobilisation:1 'immobilization'
M_moteur:1 'motor'
M_direction:6 M_cours:2 'direction'
M_ancrer:2 mouiller l'ancre:1 mouiller:2 'moor; anchor'
véhicule:2 'vehicle'
transports en commun:1 'public transportation'
XPOS_NEAR_SYN
NEAR_ANT
XPOS_NEAR_SYN
CAUSES
RELATION?
M_poursuite:5 persécution:2 'pursuit'
faire une pause:1 s'arreter:4 'pause'
moteur thermique:1 'combustion engine'
vaisseau:2 M_bateau:2 avion:2 'vessel'
véhicule à moteur:1 'motor vehicle'
M_train:2 'train'
HAS_MEROPART
HAS_MEROPART
souffler:1 reprendre haleine:1 'catch one's breath'
moteur à combustion interne:1 M_moteur:2* 'internal combustion engine'
M_boussole:1* 'compass'
M_bateau:1 'boat'

Figure 1: The MOTION source domain with corpus-specific highlights.

Dotted lines stand for relations that we added. These were labeled using EWN relation types (Vossen, 1999), where possible. As obvious from Figure 1, the domain graph would be separate partitions without our additional relations, especially those of the INVOLVED type. Conceptual relations ("typically…") are thus a necessary addition to lexical relations ("necessarily…") in order to represent conceptual source domains.

The map representing the source domain is a result of our corpus investigation of this specific target domain corpus. The structure of the source domain is not intended to be a general representation of this domain, nor does it imply fixed domain boundaries. Rather, the network shows the elements of the source domain that mapped onto the target domain from corpus attestations. If the same source domain were to be mapped onto some other target domain, other synsets might be used. A lexico-conceptual resource encoding general information on this source domain would thus have to contain more synsets and relations than those displayed in Figure 1.

The choice of source domains as well as of certain lexical items from within a source domain has the effect of "highlighting and hiding" certain aspects of the target domain. For example, among the numerous hyponyms of the central 'move' synset {*se déplacer:1 se mouvoir:1 aller:3*}–most of which are not displayed in Figure 1–, the European Constitution corpus shows a tendency towards lexical metaphors in synsets containing the verb *traverser* - 'traverse'. This profiles the path component of the motion event. The path itself is further emphasized by lexical metaphors related to the 'move' synset by INVOLVED_LOCATION and INVOLVED_DIRECTION. Also vehicles as instruments play a role in the conceptualization, but not all vehicles have metaphorical attestations in the corpus: only *train* - 'train' and *bateau* - 'boat' are found during a cross-check. Finally, synsets referring to the contrary of ‘move’ are contained within the map of the source domain. Even the ‘motor’ (as a vehicle part) and its ‘breakdown’ (causing ‘immobilization’) are thus lexically and conceptually integrated in the MOTION domain derived from our corpus.

All these highlightings and hidings can be interpreted with respect to the situation of Europe before the referendum on its Constitution: Europe is made cognitively accessible as a multi-passenger vehicle in motion on a path, which has not yet arrived but is facing obstacles to its motion, possibly resulting in being stopped.

7 Conclusion and Outlook

A method for quickly finding large amounts of lexical metaphors and characterizing their source domains has been exemplified, given a target domain corpus. The method makes use of collocate exploration of a target domain keyword, in order to identify the most promising source domains. Over 1,000 manual annotations have been obtained and will be integrated into the Hamburg Metaphor Database. This outnumbers by far the results of previous studies filed within HMD, which originated under similar conditions but did not resort to a corpus manager.

Our method is different from automated work on metaphor recognition such as (Mason, 2004) and (Gedigian et al., 2006) in that it includes nouns as parts of speech. Implementing it in an automated system would require more sophisticated lexical-conceptual resources, representing information on concrete domains (possible source domains). In particular, the addition of lexical and conceptual links between verb and noun synsets is crucial for establishing a connected source domain graph.

Acknowledgements

Thanks to Patrick Hanks, Jana Klawitter, and three anonymous reviewers for their helpful comments. – This work was supported by a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD), granted to the second author.

References

John A. Barnden, Sheila Glasbey, Mark Lee, and Alan M. Wallington. 2002. Reasoning in metaphor understanding: The ATT-Meta approach and system. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 1188–1193, Taipei, Taiwan.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *Proceedings of the 3rd Workshop on Scalable Natural Language Understanding*, pages 41–48, New York City.

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

George Lakoff. 1987. *Women, fire, and dangerous things: what categories reveal about the mind*. University of Chicago Press, Chicago.

Mark Lee. 2006. Methodological issues in building a corpus of doctor-patient dialogues annotated for metaphor. In *Cognitive-linguistic approaches: What can we gain by computational treatment of data? A Theme Session at DGKL-06/GCLA-06*, pages 19–22, Munich, Germany.

Birte Lönneker and Carina Eilts. 2004. A current resource and future perspectives for enriching WordNets with metaphor information. In *Proceedings of the 2nd International Conference of the Global WordNet Association*, pages 157–162, Brno, Czech Republic.

James H. Martin. 1994. MetaBank: A knowledge-base of metaphoric language conventions. *Computational Intelligence*, 10(2):134–149.

Zachary J. Mason. 2004. CorMet: A computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1):23–44.

Pavel Rychlý and Pavel Smrž. 2004. Manatee, Bonito and Word Sketches for Czech. In *Proceedings of the Second International Conference on Corpus Linguistics*, pages 124–132, Saint-Petersburg.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

Achim Stein and Helmut Schmid. 1995. Etiquetage morphologique de textes français avec un arbre de décisions. *Traitement automatique des langues*, 36(1-2):23–35.

Piek Vossen. 1999. EuroWordNet General Document. Version 3. Technical report, University of Amsterdam.