What can distributional semantic models tell us about part-of relations?

François Morlane-Hondère
LIMSI-CNRS, Orsay, France
francois.morlane-hondere@limsi.fr

1 Introduction

The term *Distributional semantic models* (DSMs) refers to a family of unsupervised corpus-based approaches to semantic similarity computation. These models rely on the distributional hypothesis (Harris, 1954), which states that semantically related words tend to share many of their contexts. So, by collecting information about the contexts in which words are used in a corpus, DSMs are able to measure the distributional similarity of two words, which theoretically translates into a semantic one.

In recent years, these models have become very popular in a wide range of NLP tasks (Weeds, 2003; Baroni and Lenci, 2010), mainly because of the ever-increasing availability of textual data. Regardless of their use in NLP applications, distributional data provide precious information about words' *behaviour* and their tendency to appear in the same contexts. Yet, linguists have shown little interest in DSMs (Sahlgren, 2008). We believe that this kind of information can be relied on to empirically assess the validity of linguistic theories. Conversely, by shedding light on underlying linguistic factors that influence distributional behaviours, linguistic studies can contribute to improve our understanding of the results provided by DSMs.

This paper illustrates such a qualitative linguistic approach by investigating the presence of part-of relations among distributionally similar French words. We compare distributional data and a set of part-of relations provided by humans in a lexical network. In order to assess the nature of the part-of word pairs which can – or cannot – be found in DSMs, these words were sense-tagged using WordNet supersenses. Our results show considerable discrepancies between the representation of part-of *sense pairs* in distributional data.

2 Part-of relation and DSMs

As its name suggests, part-of relation – or *meronymy*[1] – holds between a part – the meronym – and its *whole* – the *holonym* –, like in *bed/pillow*, *armor/steel* or *ostrich/feather*. It is one of the central relations used in knowledge representation.

Automatic extraction of part-of relations has been addressed using many approaches, most of which are pattern-based (Berland and Charniak, 1999; Girju et al., 2006; Pantel and Pennacchiotti, 2006). However, the unsupervised nature of the distributional approach makes it an attractive alternative.

Studies were conducted to assess the nature of the semantic relations extracted by distributional models – using human judges (Kuroda et al., 2010), thesauri (Morlane-Hondère, 2013; Ferret, 2015) or *ad hoc* datasets (Baroni and Lenci, 2011). They showed that part-of relations are present in varying proportions among distributionally similar words. This very presence is interesting in that unlike synonymy, hypernymy or co-hyponymy, meronymy is not a similarity relation (Resnik, 1993; Budanitsky and Hirst, 2006): an ostrich is not *the same kind of thing* as a feather, neither an armor is the same *kind of thing* as steel. Following the distributional hypothesis, it is not expected that these kind of meronyms share a lot of contexts.

It appears, though, that a certain proportion of them tend to do so. For example, in Baroni and Lenci (2010)'s DSM, *player*, *pianist* and *musician* are among the ten most distributionally similar words of *orchestra*. In the following of this study, we compare the semantic properties of the meronyms which can be extracted using a distributional approach and the properties of the meronyms which cannot.

[1] Some authors make a distinction between part-of relation and meronymy (Cruse and Croft, 2004).

In Vito Pirrelli, Claudia Marzi, Marcello Ferro (eds.): *Word Structure and Word Usage*. Proceedings of the NetWordS Final Conference, Pisa, March 30-April 1, 2015, published at http://ceur-ws.org

3 Methodology and data

3.1 The part-of dataset

The first step consists in gathering a set of meronyms. Although efforts are made to provide expert-built lexical semantic resources for French (Fišer and Sagot, 2008; Pradet et al., 2014), there is currently no freely-available equivalent – in terms of quality and coverage – to WordNet (Fellbaum, 1998) or the Moby thesaurus (Ward, 2002) for French. So, we use the JeuxDeMots (JDM) lexical network (Lafourcade, 2007), which is a GWAP (*Game With A Purpose*) in which players are asked to provide words which can be in a given relation with a given word[2].

Although collaboratively-built lexical semantic resources have shown to be valuable (Gurevych and Wolf, 2010) and although a relation in JDM must be provided by two different players to be added to the network, a certain proportion of part-of relations in JDM are actually hypernymys (*sucette/bonbon* 'lollipop/candy'), synonyms (*chef/patron* 'chief/boss') or thematic associations (*océanographie/eau* 'oceanography/water'). Two possible explanations for these confusions are the lack of linguistic expertise of the players or a misunderstanding of the instruction. Erroneous relations were manually removed from the set.

One interesting characteristic of JDM part-of relations is that a considerable number of them do not fit into traditional typologies of meronymy relations. For example, topological inclusions (*cell/prisoner*), attachment relations (*ear/earring*) or ownership (*millionaire/money*) are very common among JDM part-of pairs although they are considered to be non-meronymic relations (Winston et al., 1987).

After filtering the pairs whose members do not appear in our DSM and removing most of the erroneous relations, there were 24 089 part-of pairs left in our dataset.

3.2 Sense tagging

In a previous study (Morlane-Hondère and Fabre, 2012), we manually annotated the different meronymic sub-relations – following Winston and Chaffin (1987)'s typology – in a dataset like the one described above. The idea was to test whether there is a correlation between the nature of the relation between two words and their probability of being extracted in a DSM. However, the typology has proven to be inadequate, so we chose to annotate the words instead of their relation. This is also what we do in this study. This approach is inspired by the idea that the difference between the meronymic sub-relations is due to the semantic nature of the words involved (Murphy, 2003).

The above-mentioned lack of freely-available thesauri for French led us to use WordNet to perform this task. Words of our dataset were 1) translated to English, 2) mapped to WordNet synsets and 3) linked to their translation's *supersense(s)*. Supersenses – or *lexicographer classes* – are a set of 44 coarse semantic categories used to classify WordNet's noun, verb and adjective entries[3]. Examples of the 25 noun supersenses are GROUP, LOCATION or FOOD. Supersenses were then manually disambiguated (*drawer* can both belong to the PERSON and ARTIFACT supersenses, but only the latter fits in the pair *cabinet/drawer*).

3.3 The distributional model

We use a DSM[4] generated from the frWaC corpus (Baroni et al., 2009) – a 1.6 billion words corpus of French web pages.

Words in the DSM appear at least 20 times in the corpus and in at least 5 different contexts.

Syntactic dependencies were used as contexts using the Talismane parser (Urieli, 2013). Relations taken into account in the context vectors are the subject, object and modifier relations. Prepositions and coordinating conjunctions are also included as relations (the label of the relation being the preposition or the coordinating conjuction).

The weighting of the contexts was made using the pointwise mutual information and the cosine measure was used to compute the similarity between the context vectors. The minimum similarity threshold has been set to 0.02. The total number of word pairs in the DSM is 3 674 254.

4 Results and discussion

We then measure the proportion of semantically-annotated part-of pairs – *sense pairs* – in our set which are present in the DSM. Sense pairs which occur less than 100 times in the dataset are discarded. Table 1 provides the list of the 22 re-

[2] http://www.jeuxdemots.org/

[3] http://wordnet.princeton.edu/man/lexnames.5WN.html

[4] Provided by Franck Sajous from the CLLE-ERSS laboratory.

maining sense pairs and, for each one, the ratio of part-of pairs present in the DSM. In this section, we describe the *homogeneous* sense pairs – whose semantic classes are identical – and the *heterogeneous* ones, then we provide a detailed analysis of some of the PERSON/BODY meronyms which have been extracted by the DSM.

4.1 Homogeneous sense pairs

As expected, part-of relations composed of two words of the same class are the most represented in the DSM. 84 % of the TIME/TIME part-of pairs were extracted by the DSM. This can be explained by the fact that the members of pairs like *mois*/*jour* 'month/day' both appear in contexts involving temporal prepositions like *venir*_IL Y A 'to come_SINCE', *se dérouler*_DURANT 'to take place_DURING' or *scrutin*_AVANT 'election_BEFORE'.

Likewise, the spatial dimension plays a crucial role in the extraction of meronyms (78.3 % of LOCATION/LOCATION pairs are extracted). This is due to the fact that, as for time, spatial information can be conveyed by specific prepositions. Thus, LOCATION/LOCATION meronyms' shared contexts massively involve the DANS 'IN' relation.

SUBSTANCE pairs are the third best-extracted kind of pairs. The reason why 37.6 % of them has not been extracted can be illustrated by the comparison of *acier* 'steel' and two of its meronyms, namely *fer* 'iron' – which was extracted in the DSM – and *carbone* 'carbon' – which was not extracted:

1. *acier* and *fer* both appear in contexts like *grille*_EN 'grille_COMP', *forgé*_MOD 'forged_MOD' or *lame*_DE 'blade_COMP'. Thus, they appear as materials and, moreover, as materials which are used to build the same kind of things;

2. although being a material as well, *carbone* does not appear as such in the corpus. Rather, its contexts are chemical compounds like *monoxyde*_DE 'monoxide_COMP'. It is also modified by adjectives like *inorganique*_MOD 'inorganic_MOD', which describe chemical properties of *carbone*. These two kinds of contexts are not found among *acier*'s.

So, we can see that there is a discrepancy between the contexts in which *acier* appears in the corpus and the ones in which *carbone* appears: whereas *acier* – as well as *fer* – is used as a material, the representation of *carbone* that emerges from the corpus is that of a chemical element.

holonym/meronym	%	holonym/meronym	%
TIME/TIME	84	ARTIFACT/PERSON	32.6
LOC./LOC.	78.3	ARTIFACT/ARTIFACT	31.4
SUBST./SUBST.	62.4	ARTIFACT/LOC.	24.8
OBJECT/OBJECT	61	ARTIFACT/PLANT	22.8
COMM./COMM.	53.8	ARTIFACT/SUBST.	20.4
GROUP/PERSON	52.8	OBJECT/ANIMAL	19.8
LOC./ARTIFACT	46.8	PLANT/PLANT	19.7
BODY/BODY	40.5	GROUP/ANIMAL	17.1
ANIMAL/ANIMAL	41	PERSON/ARTIFACT	16.5
ARTIFACT/COMM.	39.9	ANIMAL/BODY	9.4
ACT/ARTIFACT	35.8	PERSON/BODY	5.5

Table 1: Part-of sense pairs and their presence in the DSM.

4.2 Heterogeneous sense pairs

At the other end of the scale, part-of relations composed of two words of different classes are – also logically – the less represented in the DSM.

Part-of pairs composed of words that refer to human beings or to animals and their body parts are barely present in the DSM (although being the most frequent sense pairs in our dataset). In frWaC, PERSON words appear as subjects of action (*prendre* 'to take', *dire* 'to say') or cognitive verbs (*vouloir* 'to want', *savoir* 'to know'). They are frequently modified by nationality adjectives. Body parts do not appear in such contexts. The class of body parts was actually found to be quite heterogeneous, in that body parts' distributions in the corpus differ from persons', but not in the same way:

- organ nouns mostly appear in noun compounds to indicate the location of medical interventions (*radiographie*_DE 'x-ray_MOD') or affections (*cancer_de* 'cancer_COMP' or *lésion_de* 'injury_COMP');

- limb nouns are modified by adjectives related to location and are objects of verbs like lever 'to raise' or *étendre* 'to stretch'.

All these contexts are obviously incompatible with PERSON words.

A similar distributional discrepancy can be observed with the ANIMAL/BODY sense pair, except that animal nouns tend to appear in contexts like *élevage*_DE 'farming_COMP' or *espèce*_DE 'species_COMP'. They are also modified by size

adjectives. It is interesting to note that many animal body parts like *tête*_DE 'head_COMP', *peau*_DE 'skin_COMP' or *queue*_DE 'tail_COMP' do appear among the closest contexts of animal nouns. This means that the meronymic relation between nouns referring to animals and their body parts is not a paradigmatic one. Thus, it is reasonable to say that, in order to extract this particular relation, the use of syntagmatic patterns would be a better strategy than the use of a paradigmatic DSM.

The sense pair GROUP/PERSON also presents an interesting situation. Of all the heterogeneous sense pairs, meronymic relations belonging to this one are the most likely to be extracted by the distributional method. This can be explained by a tendency to use the GROUP entities in a metonymic way: although an army is not *the same kind of thing* as a soldier, both words share contexts like *tirer*_SUJ 'to shoot_SUBJ' or *tué*_PAR 'killed_BY'. Another reason is the transitivity of properties like nationality: *armée* 'army' and *soldat* 'soldier' are both modified by nationality adjectives because usually, members of the armed forces of a nation have to be citizens of this nation.

In the section 2, we mentioned the fact that three meronyms of *orchestra* were present among its ten most distributionally similar words in Baroni and Lenci (2010)'s DSM. In our data, the meronyms *orchestre/musicien* have also been extracted: as for *army* and *soldier*, these words share semantic features. They are related to the kind of music a musician and an orchestra can play (*classique*_MOD 'classical_MOD', *traditionnel*_MOD 'traditional_MOD' or *jazz*_DE 'jazz_MOD'), the kind of actions they perform (*interprété*_PAR 'performed_BY', *accompagné*_PAR 'accompanied_BY') or their nationality.

4.3 Focus on the PERSON/BODY sense pair

In the previous subsection, we saw that meronyms belonging to the PERSON/BODY are the least likely to be extracted with the distributional approach. In this subsection, we provide further insight into this result by examining the nature of the few PERSON/BODY meronymic pairs that were successfully extracted.

The examination of the 5.5 % of PERSON/BODY meronymic pairs that were successfully extracted is disappointing: the vast majority of the contexts shared by the meronym and the holonym are quite random. For example, the meronyms *homme/main* 'man/hand' share contexts like *nu*_MOD 'bare_MOD' or *dos*_DE 'back_COMP', which are not very informative about their relation. On the other hand (!) some shared contexts like *doigt*_DE 'finger_COMP' and *saisir*_SUJ 'to grab_SUBJ' are more informative. The fact that these specific features are shared by the meronyms indicates some kind of similarity between them: when a man grabs a rock, it is actually his hand that completes the action of grabbing, as well as a man's fingers are also his hand's fingers.

The meronyms *enfant/oeil* 'child/eye' also share some interesting contexts: both the meronym and the holonym are subjects of verbs of visual perception like *regarder* 'to look', *percevoir* 'to perceive' or *observer* 'to observe'. The metonymic interpretation is quite straightforward: although the eye is the child's *part* that allows him to look/perceive/observe, this ability is extended to the whole child.

This phenomenon partially explains why such meronyms share semantic – thus distributional – features and are more likely to be extracted with a DSM.

5 Conclusion

The main goal of this study is to shed light on the linguistic phenomena at work in DSMs. By comparing a set of sense-tagged part-of relations and a distributional model, we show that the semantic class of the meronyms has a dramatic influence on their probability to be extracted by a DSM. We also highlight the – positive – influence of metonymy in the extraction of heterogeneous meronyms.

These results show that the part-of relation is not a monolithic entity but a collection of different kinds of relations between different kinds of words which may or may not be distributionally similar.

Acknowledgments

This work was partially supported by the ANSM (French National Agency for Medicines and Health Products Safety) through the Vigi4MED project under grant #2013S060.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.

Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

Marco Baroni and Alessandro Lenci. How we BLESSed distributional semantic evaluation. *GEMS 2011*, pages 1–10, 2011.

Matthew Berland and Eugene Charniak. Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 57–64, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

Alexander Budanitsky and Graeme Hirst. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47, March 2006.

D. Alan Cruse and William Croft. *Cognitive linguistics*. Cambridge: Cambridge University Press, 2004.

Christiane Fellbaum, editor. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA; London, May 1998.

Olivier Ferret. Typing relations in distributional thesauri. In Núria Gala, Reinhard Rapp, and Gemma Bel-Enguix, editors, *Language Production, Cognition, and the Lexicon*, volume 48 of *Text, Speech and Language Technology*, pages 113–134. Springer International Publishing, 2015.

Darja Fišer and Benoît Sagot. Combining multiple resources to build reliable wordnets. In *TSD 2008 - Text Speech and Dialogue*, Brno, Czech Republic, 2008.

Roxana Girju, Adriana Badulescu, and Dan Moldovan. Automatic discovery of part-whole relations. *Comput. Linguist.*, 32(1):83–135, March 2006.

Iryna Gurevych and Elisabeth Wolf. Expert-Built and Collaboratively Constructed Lexical Semantic Resources. *Language and Linguistics Compass*, 11(4):1074–1090, 2010.

Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

Kow Kuroda, Jun'ichi Kazama, and Kentaro Torisawa. A look inside the distributionally similar terms. In *Proceedings of the Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010)*, pages 40–49, Beijing, China, August 2010. Coling 2010 Organizing Committee.

Mathieu Lafourcade. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07: 7th International Symposium on Natural Language Processing*, page 7, Pattaya, Chonburi, Thailand, December 2007.

François Morlane-Hondère. *Une approche linguistique de l'évaluation des ressources extraites par analyse distributionnelle automatique*. PhD thesis, Université de Toulouse II le Mirail, 2013.

François Morlane-Hondère and Cécile Fabre. Étude des manifestations de la relation de méronymie dans une ressource distributionnelle. In *Proceedings of TALN 2012*, Grenoble, France, June 2012.

Lynne Murphy. *Semantic Relations and the Lexicon*. Cambridge University Press, New York, 2003.

Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 113–120, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Quentin Pradet, Gaël de Chalendar and Jeanne Baguenier Desormeaux. WoNeF, an improved, expanded and evaluated automatic French translation of WordNet. In *GWC 2014*, Tartu, Estonia, 2014.

Philip Resnik. *Selection and Information: a Class-Based Approach to Lexical Relationships*. PhD thesis, The Institute For Research In Cognitive Science, University of Pennsylvania, 1993.

Magnus Sahlgren. The distributional hypothesis. *Rivista di Linguistica*, 20(1):33–53, 2008.

Assaf Urieli. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université de Toulouse II le Mirail, 2013.

Grady Ward. *Moby Thesaurus List (English),*. 2002.

Julie Weeds. *Measures and Applications of Lexical Distributional Similarity*. PhD thesis, Department of Informatics, University of Sussex, 2003.

M. E. Winston, R. Chaffin, and D. Herrmann. A taxonomy of part-whole relations. *Cognitive Science*, 11(4):417–444, December 1987.