EPiC Series in Language and Linguistics
Volume 1, 2016, Pages 45–57
CILC2016. 8th International Conference on Corpus Linguistics

The meaning of hashtags matters: detecting hashtags such as *#JeSuisKouachi*

Daniel Couto-Vale[1] and Adjan Hansen-Ampah[2]

[1] IfAAR - English Linguistics, RWTH Aachen University
danielvale@icloud.com
[2] IfAAR - English Linguistics, RWTH Aachen University
adjan.hansen@ifaar.rwth-aachen.de

Abstract

In this paper, we model how the contexts of situation and of discourse restrict meaning potential and apply the resulting model in automatic linguistic analysis of hashtags for tasks of sentiment analysis. With our model, we are able to assign different meanings to *#JeSuisCharlie* and *#JeSuisKouachi* and variations of them.

1 Introduction

Following the Paris terrorist attacks in January 2015, a French man held a sign reading *Je Suis Kouachi* (I am Kouachi) at the Republican March in memory of the victims. Since Kouachi and Coulibaly are the surnames of the terrorists, he was taken under arrest and charged 3 months detention and €1,000.00 euro fine for provocation and threat of violence [11]. The same man had signs reading *Je Suis La Vie* (I am life), *Je Suis Humain* (I am human) and *Je Suis Charlie* (I am Charlie) and he was not charged for those other statements. This is strong evidence that these clauses – though compositionally similar – have different discursive effects.

Understanding the differences between these hashtags is relevant not only for law enforcement, but also for those monitoring hate speech on the web. As Oboler [9], a member of the Online Hate Prevention Institute, puts it, "In social media a variety of hashtags went viral from *Je Suis Charlie*, the French Muslim's response *Je Suis Ahmed*, solidarity with the Jewish community through *Je Suis Juif*, and the inclusive German response of *Je Suis Humain*. While all mourn the deaths of those killed, different values led to different messages gaining traction on social media in different communities. There is also a warning sign in the rise of the *Je Suis Kouachi* hashtag."

Therein we can notice that both the police and Oboler distinguish *Je Suis Charlie*, *Je Suis Humain*, *Je Suis Juif* from *Je Suis Kouachi* and the rationale they used for this distinction does not seem to be based on grammatical composition alone. For instance, the police did not charge the man holding signs for impersonating Charlie or one of the Kouachi brothers, and Oboler understood that Jews and Non-Jews alike demonstrated solidarity with the Jewish community through *Je Suis Juif* (I Am A Jew).

A.Moreno Ortiz and C.Pérez-Hernández (eds.), CILC2016 (EPiC Series in Language and Linguistics, vol. 1), pp. 45–57

Figure 1: Charlemagne statue in front of the city parliament house in Aachen

In addition, there is also linguistic evidence that these clauses are not only semantically but also grammatically different from more usual uses of *Je Suis* (I am). In the days after the attacks, tweeters also posted *#JeSuisCharlieMaisPasTrop* (I am Charlie but not very much). In this hashtag, the presence of an intensifier such as *PasTrop* is a grammatical cue that the represented relation between two entities is gradable. This is a very different kind of meaning from that of a statement by a person named Mike such as *je suis Mike* or *je m'appelle Mike* (my name is Mike). A person named Mike would not be able to tell his listener *je suis Mike mais pas trop* (I'm Mike but not very much) or *je m'appelle Mike mais pas trop* (my name is Mike but not very much). A rationale for this difference is that the meanings of having solidarity with victims and of supporting terrorists can have different degrees of intensity while the relation between people and their own names cannot.

Yet, the difference between *#JeSuisCharlie* and *#JeSuisMike* does not separate the two grammatical structures completely. Even if saying one's own name after *#JeSuis* is different from saying someone else's name, at least in the rush of the moment, when a tweeter named Mike concluded a tweet about the attacks with *#JeSuisMike*, noticing the replacement of *Charlie* by *Mike* might be crucial for understanding what he meant, which may have been *#JeSuisPasCharlie* (I am not Charlie). The same effect might not have been so easily achieved if Mike had posted *#JeMAppelleMike* because that is not immediately noticeable as similar to the trending hashtag *#JeSuisCharlie*. The same double reading seems to be demanded from readers in Figure 1 [8]. There the statue of Charlemagne (Charles the Great) is decorated with a sign reading *Je Suis Charlie*. In this case, *Charlie* might have been understood as the endearment form of the name *Charles*. And the message might have been understood as a statement by whoever placed the sign to be about a personal name together with an endearment of Charlemagne. However, in the days after the attacks, it might also have been understood as a statement by that person that Charlemagne would have solidarity with the victims if he were alive. In other words, the same clause supports two statements simultaneously and the pun in this figure only works if those two readings occur.

In this context, the question that we raise is how a shared collective experience, including what has been said, restricts the kinds of linguistic meanings that can be construed and guide our understanding. In the following, we shall model this restriction of meaning potential and use the resulting model for the automation of linguistic analysis of Twitter hashtags, which can be applied in law enforcement, in monitoring of hate speech on the web and in other tasks of

sentiment analysis. Due to space constraint, we leave for future work the task of modelling the discursive effects of grammatically alternative meanings as in the case of the tweet ending with *Je Suis Mike* instead of *Je Suis Charlie* and multiple meanings as in the sign *Je suis Charlie* on the statue of Charles the Great.

2 Experience and Discourse

Before modelling meaning potential and the restriction thereof, it is important to make clear what we mean by 'experience' and 'discourse'.

We conceive of experience as descriptive models of realities that we humans accumulate throughout time. As a model, experience contains entities and relations, the types of these entities and the types of these relations. It also includes static images and dynamic image flows that can be cut into entities in various ways [12].

We conceive of discourse as a sequence of entities associated with a sequence of clauses. On the one hand, a primary clause may represent a primary figure of someone doing something and a secondary clause may represent a secondary figure of someone doing something else. In such cases, the secondary figure is taken to expand the primary because the discourse episode grows with the advent of the secondary figure. On the other hand, a clause may alternatively construe a figure where someone says or thinks about something. In doing so, it may embed a secondary discourse into the primary one, what is called projection [4, p. 443]. At every level of projection, a discourse may be completely or predominantly monologic, dialogic, or multilogic depending on the number of sayers or thinkers (signers) that participate in the discourse construction.

In such an approach, a discourse is conceived of as an unfolding sequence of configured or mentioned instances of classes of experience entities. A set of equivalent discourse entities (equivalent tokens) corresponds to a single experience entity (the same value). This means that, if English speakers start talking about *a dragon* and mention *the dragon* several times, each mention of the dragon construes a different discourse entity, but all these discourse entities are equivalent and together they construe one and the same experience entity, namely the dragon.

In addition to discourse, there are perception and simulation. In the same way as discourse entities, a set of equivalent perception entities and simulation entities may also construe a single experience entity. Experience is a super-stratum populated by entities (values) that correspond to multiple entities (multiple tokens) in discourse, perception, and simulation.

Linguists can talk both of personal or (distributed or shared) collective experience. When we talk about collective experience, we must be aware that "collective" implies a group of people and that "shared" implies a subgroup of those people. There are, therefore, as many distributed collective experiences as we can make groups out of people and as many shared collective experiences within a group as we can find subgroups in that group. In this sense, whenever we talk about a collective experience, we must specify the group that has the experience and whether this experience is distributed or shared. And if it is shared, we must specify the subgroup that shares the experience.

In addition, extending the dragon example into perception, if an English speaker is at an amusement park and gets asked *have you seen the dragon?*, the meaning of *the dragon* is an instance of dragon in the discourse, i.e. a discourse entity that is a member of the class of dragons. In turn, this discourse entity construes an experience entity, which is also an instance of the class of dragons. If a few minutes later the English speaker sees a large statue of a dragon, the perceptual meaning of this statue is also an instance of dragon, that is, a perception entity that is a member of the class of dragons, which also construes an experience entity. At this point, the English speaker might conclude that the dragon that one is expected to see is the one

represented by the statue. And in that moment, the perception entity is taken to be equivalent to the discourse entity and to construe together with it one and the same experience entity. In other words, a dragon does not need to exist currently as a tangible body in our physical world, i.e. as an entity cut out of the same reality of which we are also parts, to be taken as an experience entity, a discourse entity, a perception entity or an simulation entity.

By conceiving of experience in such a way, we make no claim that what is experienced is known. In other words, the experience of there being a dragon in the amusement park is not to be confounded with knowledge that there is a dragon statue there: the experience of there being a dragon does not imply that there is a dragon whereas the knowledge that there is a dragon statue implies that there is a dragon statue. Experience and knowledge are not the same kind of abstraction. Moreover, the experiential meaning of *ein Drache*, *un dragon*, *a dragon* is such that for a sign **to mean** a dragon is for it **to construe** a dragon in discourse and for that dragon on its turn to construe a dragon in experience. Likewise, a statue that has the shape of a dragon does not need to be a dragon itself for it to construe a dragon in a viewer's perception and experience. A dragon statue means a dragon perceptually. And since viewers know the statue is not a dragon itself, they know there is no dragon in the amusement park even though they are capable of perceiving a dragon and construing an experience of it being there.

Finally, from an empirical standpoint, we can talk about classes of linguistic form and classes of perceptual forms (language and perception). We can talk about an experiential meaning in terms of a base of methods for construing entities in experience and about an experience entity base in terms of a base of experience entities. This understanding of experience as a super-stratum is not equivalent to an understanding of knowledge in cognitive computing as in "knowledge base" according to which a noumenon (a known thing) must correspond to a perceived/perceivable referent "out there" where the knower is. It is also not equivalent to a bichotomous model of semiotics: the signifier (signifiant) and the signified (signifié) in Saussure's terms or the name and the named referent (Bedeutung) in Frege's terms; nor is it equivalent to a trichotomous model of semiotics: the sign (the signifier, the name), the idea (the signified), and the referent (the named, the known, the perceived) in Peirce's terms. In this understanding, reference is also not the relation between a set of co-referential discourse entities and a part of what is 'out there' called referent but the relation between a set of equivalent entities in discourse, perception, and simulation (the tokens) and an experience entity (the value).

In short, recognised forms are understood as standing for signs that stand for a semantic token. Experientially equivalent semantic tokens correspond to the same experiential value. And this stratification works for discourse, perception and simulation alike. A person's experience is that which the person has experienced so far and a group's experience is that which a group has experienced so far. When talking about a group's experience, we can either consider the whole experience that is distributed among the members, including parts of it that not all members have in their own experiences (the distributed collective), or we can consider only the overlap between the experiences of interaction participants (the shared collective experience).

2.1 Names and class names

When parsing hashtags, **thing names** were taken to be any lexical item that 'redounded' [2, 6] with a finite set of phenomena that we assumed to be equivalent, that is, that we assumed to construe the same experiential thing [3, p. 73]. Figure 1 represents a series of perception snapshots in each of which there is an image of a human. These images can be experientially assumed to be the same human. That human can be said to have a name as in *this is Charlie* and the same human can be mentioned by name as in *Charlie is standing there*. All such

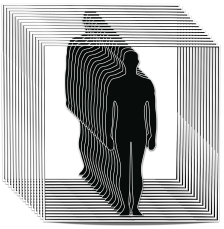

Figure 2: Thing

mentions of that human by name were taken to be mentions of the same experiential thing. This means that experiential entities of type 'thing' can have names and that thing names can be used to mention them.

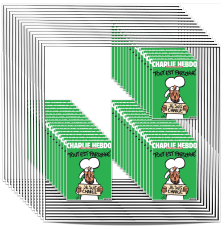

Figure 3: Thing Class

In contrast, **thing class names** were taken to be lexical items that redounded with a set of non-equivalent phenomena, that is, of phenomena that were not taken to be the same experiential thing. Figure 3 shows a series of perception snapshots in each of which three magazines are to be seen. A magazine of that set may be said to be an instance of a class as in *this is a Charlie* and the same magazine may be mentioned as an instance of this class as in *this Charlie is mine*.

Moreover, **process class names** were taken to be lexical items that represented 1) a relation between things and their attributes such as *is* in *this is Charlie* and *is* in *this is a Charlie*, 2) a relation between identical discursive and perceptual tokens such as *is* in *Charlie is that man*, 3) relations between two different things such as *have* in *I have a Charlie*, 4) mental processes such as *am* in *I am Charlie*, 5) saying processes such as *ask* in *ask Dagi*, and 6) material processes such as *search* in *don't search for me*.

3 Hashtags

For this paper, we collected a corpus of hashtags that were tweeted between 9-15 of January 2015, comprising the days following the terrorist attacks in Paris, 7-9 of January 2015.

We filtered tweets containing one or more forms out of the 180[1] most frequent word forms in German, French, and English. We used the 60 most frequent word forms found in the German Corpus of IDS [5], in the French Corpus of MENESR [7], and in the COCA Corpus [1]. We compiled a corpus with a total of 205,003 hashtags (form tokens), that instantiate 49,249 different form types.

[1]Twitter imposes a limit of 180 searches per timeframe of 15 minutes.

Since we are interested in compositionally similar clauses that have different discursive effects due to shared collective experience, we only looked at hashtags that are comprised of more than one word and that have at least a group of words that represents a named entity such as *Charlie* in *#JeSuisCharlie* or an instance of a named class of things such as *Juif* in *#JeSuisJuif*.

3.1 Discourse semantics

We conceive of the grammatical unit of clause as the whole sequence of words that represents a figure in discourse (see Chapter 2). This notion of clause is not equivalent to the notion of sentence in a generative model of language. For instance, *#cequejaimedanslesquatrequarts* (that's what I like about cup cakes) represents a single process of liking something about cakes, therefore it is taken to be a single clause and not as a sentence with an embedded sentence that has a cleft structure interpretation. As far as experience is concerned, the constituents of the clause were taken to be groups of words that correspond to entities in discourse such as things and processes [4]. This means that grammatical structures – as we modelled them – are associated with semantic structures in discourse (henceforth *rhetoricosemantic* structures[2]).

A figure was taken to have a slot for a process – what is going on – and a slot for each element that participates in this process. The circumstances in which the figure is to become perceivable (prototypically a spatiotemporal feature or condition) was not taken to be a complement of the figure (a slot-filling element) but an extension to it (an appended element). Grammatically, semantic complements were treated either as being represented by a constituent of the clause such as the mention of *Trick* in *#FollowTrick* or as being realised by a feature of the Finite constituent of the clause such as the addressed subject of *#FollowTrick*.

Not all clause constituents had experiential functions in a figure: process, participants, and circumstances. For instance, personal names such as *Taylor* in *#FollowMeTaylor* were used to determine the addressee of an imperative clause whenever tweets were addressed to that named person. In such cases, the figure is completely represented by the other elements of the clause – e.g. *FollowMe* in *#FollowMeTaylor* – and the addressed person fills the addressee slot in an interpersonal structure called address. The words *I* and *you* were interpreted as mentions of whoever fills the slot of, respectively, addresser and addressee in this parallel structure.

Within such a model, the same form can be associated simultaneously to one or multiple words. For instance, *#jesuismini* as one word is the nickname of a woman. At the same time, this hashtag can be understood as a clause meaning 'I am small'. Both understandings are valid and may co-occur in any instance of this form.

3.2 Automation

From this perspective, one way to explain what tweeters do is to conceive of word forms as something that gets projected onto the hashtag. With such an explanation, it would be the limits of matching segments that would count as word boundaries, and not spaces. In theoretical terms, since segments would match word forms independently of whether there are non-letter characters representing word boundaries, one can say that this kind of form recognition is lexicogrammatically motivated, i.e. not independent of lexis and grammar, and ultimately rhetoricosemantically motivated, i.e. not independent of discourse semantics.

In automatic linguistic analysis, the task of creating a word sequence (wording) for *#jesuis-charlie* was performed in the following way. First, a form recogniser identifies all segments of a character sequence that can instantiate a word form and creates a chart such as the one in Figure

[2]Taking a discourse to be a rhetorical structure.

j	e	s	u	i	s	c	h	a	r	l	i	e
1-1	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11	1-12	1-13
	2-2	2-3	2-4	2-5	2-6	2-7	2-8	2-9	2-10	2-11	2-12	2-13
		3-3	3-4	3-5	3-6	3-7	3-8	3-9	3-10	3-11	3-12	3-13
			4-4	4-5	4-6	4-7	4-8	4-9	4-10	4-11	4-12	4-13
				5-5	5-6	5-7	5-8	5-9	5-10	5-11	5-12	5-13
					6-6	6-7	6-8	6-9	6-10	6-11	6-12	6-13
						7-7	7-8	7-9	7-10	7-11	7-12	7-13
							8-8	8-9	8-10	8-11	8-12	8-13
								9-9	9-10	9-11	9-12	9-13
									10-10	10-11	10-12	10-13
										11-11	11-12	11-13
											12-12	12-13
												13-13

Figure 4: Form Recognition

4. This chart contains overlapping word forms as in [1: [*j*, *je*], 2: [*es*], 3: [*suis*], 7: [*charlie*]]. Secondly, a parser using the Cocke-Younger-Kasami parsing algorithm (CYK algorithm) and having characters as atoms can produce composite grammatical structures with a partially filled chart such as the one above and finish filling the chart with the composites. In particular, we used our own customised version of the OpenCCG api (available at https://goo.gl/YPtmp0) that is able to recognise word forms and cope with the size of a character chart. Finally, a separate process recognises that this sequence of characters can realise a meaningful sequence of words [*je*, *suis*, *charlie*].

This process bypasses tokenisation ('word segmentation' or 'word breaking') in the sense of string partitioning and pos tagging since there is no actual segmentation of strings into substrings (grams) until the parsing process ends. In addition, this process is able to give two different segmentations to a hashtag such as *#jesuismini*: one of which is a mention or an addressing of a woman by her nickname 'Jesuismini' and the other a clause representing that the addresser is small or is called 'Mini'. Such an understanding that is based on two parallel segmentations of a string cannot be achieved if string partitioning is realised before parsing.

3.3 Dealing with a hashtag corpus

The frequency of hashtags takes the form of a long tail curve. This means that there are a small number of hashtags that happen very frequently and a large number of hashtags that happen once or twice. Therefore, a gold standard corpus of random hashtags would include a myriad of infrequent and single-occurrence nicknames and class names that would not be useful for the automatic analysis of new data, especially so for the purpose of law enforcement and monitoring of hate speech.

For this reason, we opted to tackle the issue of modelling language in a more cost-efficient way. We created a sub-corpus of frequent hashtags and manually annotated it. This sub-corpus includes frequently used names of things and of classes of things that do also occur in multi-word hashtags in the long tail. With this corpus, we are able to build a linguistic model that is useful for retrieving long-tail multi-word hashtags that are composed of annotated words in the sub-corpus. Because we retrieve less than one hashtag per thousand from a long-tail, we cannot provide a value for recall. And because the result of an automated linguistic analysis are merely alternative possible meanings, we cannot account for precision. We can only recognise possible clauses and possible figures for a particular character sequence with the current approach. It would be the task of a law enforcer or an institute monitoring hate speech to judge whether

the instances of meaning they are interested in are plausible in the context of situation given the discourse it is embedded in. The ability to detect such instances in thousands would be the benefit they would have when using this tool.

We created a sub-corpus of hashtags with more than one word. Due to time constraints for the annotation task, the sub-corpus includes only two kinds of multi-word hashtags: those that occurred at least 50 times and those that occurred at least 3 times and contained a frequent entity name or a frequent entity class name. By "frequent", we mean those that occurred alone as a one-word hashtag at least 100 times. This was a way of limiting the number of hashtags so that we could manually annotate them with the available working hours.

4 Linguistic Modelling

Our annotation schemes were developed with the UAMCorpusTool [10]. They consist of regions within a larger systemic network.

4.1 Language Typology

We created a region for language typology in our systemic network. We interpreted thing names such as *#CharlieHebdo* and *#Pegida* to have a meaning that was not language-dependent. For this reason, we included a system of UNIVERSALITY with two features to distinguish *language-independent* hashtags from *language-dependent* ones. In addition, since we were interested in analysing the hashtag corpus in three languages, we included a system of LANGUAGE with *german*, *french*, *english* and *other-languages* as features (see Figure 5). This typology allowed us to process hashtags in one language even when it is part of a tweet in another language. This means we did not recognise the language before parsing, but we parsed the hashtags and the presence of a wording in the filled chart for a given language was taken as evidence that the hashtag was an instance of that language.

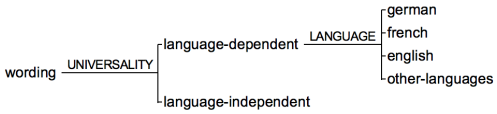

Figure 5: Language typology region

4.2 Graphology

We were able to systematise graphological variation motivated by word type, sentence[3] boundaries and word form boundaries with five systems (see Figure 6). Whether word forms in a tweet are finalised by a space and/or start with a capital letter is both language-specific and register-specific.

When it came to hashtags, we found a variety of capital standards as reactances to the same lexicogrammatical structure within the same language: e.g. *#JeSuisCharlie*, *#JesuisCharlie*, *#jesuisCharlie*, and *#jesuischarlie*, *#Je_Suis_Charlie*, *#Je_suis_Charlie*, *#je_suis_Charlie* and *#je_suis_charlie*. In such cases, a word form might be written *with-initial-letter-capital* not only for realising word features or sentence and sub-sentence boundaries as outside of hashtags, but also for realising word form boundaries such as the capital letter *S* in *JeSuisCharlie*.

[3] In a systemic functional theory, a sentence is a bounded sequence of word forms.

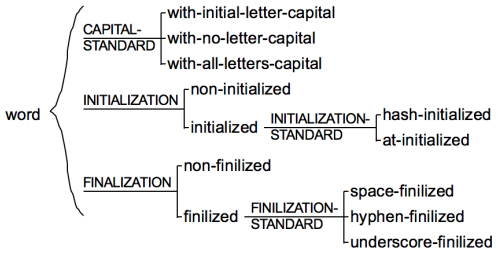

Figure 6: Graphology region

4.3 CCG Categories

We modelled grammatical structures as instances of grammatical categories using a combinatory categorial theory of language [13]. The ranks of lexicogrammatical structure were clause (c), phrase, groups, and words as in Systemic Functional Linguistics (SFL) [4]. Groups include nominal groups, groups of words that represent things. We shall call nominal groups mentions of things (m) because, as far as this paper is concerned, the notion of nominal groups can be seen as related to the more well known notion of 'mentions' outside of Systemic Functional Linguistics in the research communities of Mention Detection (MD) and Mention Head Detection (MHD)[4]. Ranks of lexicogrammatical structure are analogically patterned to graphological structures such as sentences, subsentences and letter sequences (grams as in n-grams) [4, p. 16][5]. Ranked structures should be conceived of as being associated to rhetoricosemantic structures as explained in Section 3.1.

For mentioning things, several lexical items were used. For non-classified non-named things, we saw deictics (m) such as *Das*, *Ça*, *CeLa*, *This*, and *ThatThing*. For named things (persons included), thing names (m) such as *Taylor*, *CharlieHebdo*, and *JeSuisMini*[6] have often been used. For classified things, thing class names (tc) such as *Shit* in *ThatShit*, *Charlie* in *MonCharlie* and *NonMuslims* in *AllNonMuslims* occur together with a deictic (m/tc) such as *that*, *mon* and *all*.

Grammatical categories for multi-word lexical items have been given two kinds of grammatical complements: on the one hand, there are grammatical structures that represent complements of figures; on the other hand, there are grammatical structures such as *Hebdo* in *#CharlieHebdo*, *La* in *CeLa*, and *A* in *TouchePasAMonCharlie* that are fragments of a multi-word lexical item. Such fragments do not fill slots of nor append elements to the figure and the address. They only complete the grammatical structure that includes other fragments of the lexical item.

Other categories exist in our linguistic resources but are not reported here due to limited space.

[4] We philosophically align with SFL when it comes to a phenomenological and consensual take on experience construction and we do not take mentions of things to be knowledge representation as is typical in MD and MHD.

[5] Lexicogrammatical and graphological structures overlap with syntactic units of analysis such as sentences (s), prepositional phrase (pp), and noun phrase (np), but, since they are teleological/functional classes of grammatical structure, they are not equivalent to their approximate non-functional counterparts.

[6] *JeSuisMINI* is the pseudonym of a female tweeter.

4.4 Lexicogrammar

We performed a systemic functional transitivity analysis of all clauses in our hashtag sub-corpus [3, 4]. We were able to distinguish 22 figure types for major and minor clauses.

With the Finite/Predicator categories c/m, we found imperative clauses representing transitive material processes that alter the social graph[7] as in *#FollowTrick*, *#FollowMeTaylor* and *#FollowMeCaniff* (where Taylor Caniff is the addressee) and addressed verbal processes as in *#AskAustin*. Imperative clauses representing transitive processes need to fill two semantic arguments: an addressed subject and a mentioned object.

For the same grammatical category, there were volitive clauses representing intransitive material processes such as *#RIPCaroline* (where Caroline is a mentioned non-addressed person who has passed away). The only argument of such volitive clauses was the mentioned subject, which happened after the Finite/Predicator.

Some instances of *#FollowMeTaylor* were addressed not to Taylor but to other fans of his. In such cases, since these tweeters did not have the person mentioned as an addressee, such clauses had the nature of a wish and Taylor became a mentioned non-addressed subject of the clause. The Finite/Predicator category for volitive clauses representing transitive material processes in English was $c/m/m$.

For categories $c\backslash m/m$, we found declarative clauses representing emotive mental processes: solidarity as in *#JeSuisCharlie* and *#JeSuisAhmed*, support as in *#JeSuisCharlie*, *#JeSuisDieudonné*, and *#JeSuisKouachi* (see Section 4.5 for contextual restrictions), and devotion as in *#JAimeMonProfete*. There were also possessive relations of ownership such as *#JAiMonCharlie* (I got a Charlie).

For the category $c\backslash m/tc$, we found a series of declarative clauses representing relational processes. There were ascriptions of religious identity *#JeSuisMuçulmain* and ethnic identity *#JeSuisNigerian* amongst others. The same grammatical category also represented solidarity with victims that have those identities.

In French, we gave special treatment to the word *Mon* in both *#JAiMonCharlie* (I got a Charlie) and *#TouchePasAMonCharlie* (don't mess with Charlie). In the first case, *mon* was taken to be a reflexive deictic that has the category m/tc and to be a preselected feature of the mention for the lexical item *avoir son*. It was not understood as a possessive deictic since there can be no **#JAiTonCharlie* and no **#JAiLeCharlieDeMonAmi* with similar rhetoricosemantic structures. In the second case, the *mon* was not taken as a deictic at all and was given the category m/tn. We understood it as representing a protection relation between the addresser and the named entity. The motivation for this protection relation seems to have been what the named entity Charlie stood for, namely *#LibertéDExpression* (freedom of speech), which was a frequently cooccurring hashtag.

4.5 Taxonomy

We noticed that topical entities in a discourse restrict meaning potential in at least two ways in our hashtag corpus. The first is that less delicate classes of things are used for a topical thing whereas more delicate classes of things are used for non-topical things. For instance, once "junk food" becomes the topic, it can be mentioned with a named class of things such as *Shit* in *#ILoveThatShit*. The second way in which it changes meaning potential is through a situational taxonomy. Every figure in the discourse can be interpreted as a taxonomical classification of

[7] For future record, tweeters can currently alter the social graph by adding and removing other tweeters to and from a personal list of those that they want to receive tweets from. These two material actions are respectively *to follow someone* and *to unfollow someone*.

discourse entities. For instance, if the discourse includes a figure in which terrorists attacked Charlie, the entity named Charlie becomes an instance of an affected targeted thing, i.e. a participant of an affecting transitive process. More delicately, it also becomes an attacked thing, what includes not only the participation role but the kind of process that affected it. Yet more delicately, it becomes a thing attacked by terrorists, a thing class that includes not only the named class of process but also the named class of the other participant (see Figure 7).

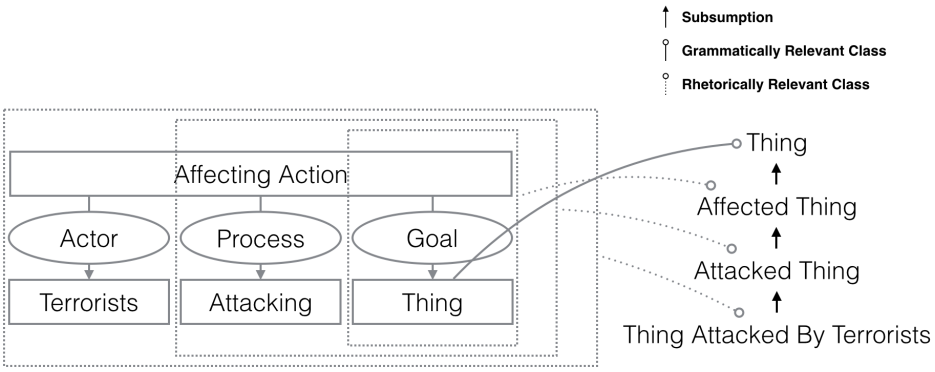

Figure 7: Taxonomy based on a figure in discourse

This has the effect that the meaning of *#JeSuisCharlie* as an emotive mental process of solidarity is only possible when the discourse includes the attacks, thus having Charlie as an attacked thing. This has the additional effect that, if the discourse includes only a figure in which Charlie offends all muslims, the meaning potential of *#JeSuisCharlie* may include an emotive mental process of support for the offence. And in case both figures are present, clauses such as *#JeSuisCharlieMaisPasTrop* (I'm Charlie but not very much) become possible since the intensifier may function as a cue that the tweeter subscribes to a set of interpretations and not to others.

4.6 Application

When specifying figures for process classes, we were able to make restrictions to clause and figure complements: respectively a lexicogrammatical signature and a rhetoricosemantic signature for each figure type. Due to space constraints, we shall discuss restrictions only to mentions of things and to mentioned things. At the grammatical stratum, mentions have been filtered by marker/case: for instance, the French a-marker for *AMonCharlie* in *#ToucheAMonCharlie* and the German accusative case for *Mich* in *#FragMichNicht*. At the rhetoricosemantic stratum, mentioned entities have been filtered by entity class: for instance, in one interpretation procedure, the entity named Charlie was taken to be an entity that had been attacked by terrorists, an attacked entity, an affected entity. With such a rhetorical taxonomy, we were able to make only affected entities be accepted as an argument for the solidarity figure as in *#JeSuisAhmed* and *#JeSuisCharlie*. Similarly, we made only acting entities be accepted as an argument for the support figure as in *#JeSuisKouachi* and *#JeSuisCharlie*. As a result, when no action of Charlie was present in the discourse, the analyser was able to construct the figure of solidarity while filtering out the figure of support; and when an action was present, the analyser was able to construct both figures of solidarity and of support.

5 Conclusion

As a product of our work, we have a software kit that performs a linguistic analysis of hashtags from any web platform (Twitter, Facebook, Youtube...) at the stratum of discourse semantics, given that named things and classified things are taxonomically classified for a given context of situation. We warn that such rhetoricosemantic taxonomies are neither stable nor uniform among tweeters. This is always the case when we are dealing with a large linguistic community since different individuals may have very divergent ideas of what is widely known and in no need of being "restated". In addition, tweets are very short and oftentimes they do not provide enough information for the construction or selection of a taxonomy for discourse entities. In addition, the presumed discourse includes what has been said prior to the tweet, typically what has been published on popular information channels (TV, press, online), but not restricted to that. That said, it is not possible to analyse a hashtag in a rhetoricosemantically motivated way 'solely' based one the text of the tweet. Shared and distributed experience must also be taken into account. However, given that this paper concentrates in modelling how a shared collective experience restricts the kinds of linguistic meaning that can be construed, judging what particular individuals consider to be the shared collective experience goes beyond our scope.

Acknowledgments

The work on the Parser was performed the German Research Foundation (DFG) project TRICKLET (Translation Research in Corpora, Keystroke Logging and Eye Tracking), research grant no. NE1822/2-1. The linguistic annotation of hashtags was performed within the IfAAR – English Linguistics.

References

[1] Mark Davies. Word frequency data. Technical report, Brigham Young University, Provo, Utah, July 2012.

[2] Michael A K Halliday. Language and the order of nature (1987). In Jonathan J Webster, editor, *On Language and Linguistics*, pages 116–138. 2003.

[3] Michael A.K. Halliday and Christian M.I.M. Matthiessen. *Construing experience through meaning: a language-based approach to cognition*. Continuum, London/New York, 1999.

[4] Michael A.K. Halliday and Christian M.I.M. Matthiessen. *An Introduction to Functional Grammar*. Oxford University Press, New York, 2004.

[5] Institut für Deutsche Sprache. Korpusbasierte Wortformenliste. Technical report, Institut für Deutsche Sprache, Mannheim, April 2009.

[6] J Lemke. Redundancy and Morphogenesis: coding-identifying relations and their relevance to social positioning and bureaucratic processes. pages 1–66. April 2007.

[7] MENESR. Liste des mots classée par fréquence décroissante. Technical report, Ministére de l'Éducation Nationale, de l'Esseignement Supérieur et de la Recherche, Paris, 2014.

[8] Sebi Meyer. Charlemagnhe statue in front of the city parliament house in Aachen officially decorated with a sign reading "Je Suis Charlie" in the days after the terrorist attacks of Charlie Hebdo, January 2015.

[9] Andre Oboler. Je Suis Humain: Responsible free speech in the shadow of the Charlie Hebdo murders. Technical report, Online Hate Prevention Institute (OHPI), 2015.

[10] Michael O'Donnell. UAM CorpusTool: Guia do Usuário Versão 2.6. Technical report, Wagsoft, Madrid, December 2010.

[11] Ouest-France. Justice. Lors du rassemblement, il brandit « Je suis Kouachi ». Technical report, Ouest-France, March 2015.

[12] Barry Smith and Berit Brogaard. A Unified Theory of Truth and Reference. *Logique et Analyse*, 43:1–46, 2003.

[13] Mark Steedman and Jason Baldridge. Combinatory Categorial Grammar. In Robert Borsley and Kersti Borjars, editors, *Non-Transformational Syntax*, pages 181–224. Oxford UK, 2011.