ANNOTATING ARGUMENT DROP IN THE SWISS WHATSAPP CORPUS

Franziska Stuntebeck (franziska.stuntebeck@uzh.ch)

1. INTRODUCTION

WhatsApp messages are a precious source for syntactic and pragmatic studies on non-standard writing. Nevertheless, grammatical research on the linguistic properties of WhatsApp messages is exceedingly rare (cf. Ueberwasser/Stark 2017). Although argument drop has been described as a typical register marker of written abbreviated or simplified registers (such as diaries: cf. i.e. Ihsane 1998, Haegeman 2017; recipes: cf. Massam/Roberge 1989, Weir 2017; notes: cf. Janda 1985), as well as for different forms of electronic writing (cf. Panckhurst 2009, Fairon et al. 2006, among others), it has, to my knowledge, not been systematically studied so far. Only the omission of subjects has been analyzed thoroughly for French and German text messages in the doctoral theses of Robert-Tissot (2018, see also Stark/Robert-Tissot 2017) and Frick (2017)[1], as well as for English and French diaries by Haegeman (2017, among others). Stark/Meier (2018) conducted a pilot-study on argument drop in French and German WhatsApp messages. Electronic chats like WhatsApp messages provide an ideal set of written interactive data not only for the study of syntactic properties of null arguments, but also for identifying their information-structural properties (e.g. D-linking), thanks to the preceding context of the elements in question provided by the chat. This is shown in (1), where a given sentence topic is dropped in a French WhatsApp message:

(1) J'ai commencé un ppx$_i$ […] Je t'envoie ça$_i$ ajd si j'Ø$_i$ ai fini.[2]
I have begun a ppt […] I you send that today if I have finished.
'I started a PowerPoint presentation […] I will send that to you today if I have finished [it].'

With my research project, I aim to fill this gap of systematic studies on WhatsApp messages. The overall research question is to discover whether argument drop is technically motivated or a register-specific feature (namely if there is, or is not, a difference between the older text messages and newer WhatsApp messages, usually written on a virtual keyboard and smartphones). Moreover, I examine whether argument drop in non-standard electronic writing is an instance of *topic* drop − as argued by Robert-Tissot (2018) − or if it is better analyzed as a truncated structure − as proposed by Haegeman (2013, among others) (cf. section 2), or neither of these. Hence, the empirical analysis is based on the syntactic constraints that have been identified for written subject omission in English and French diaries (cf. Haegeman 1997, among others) on the one hand, and for the *familiar topic* drop analysis in French text messages (cf. Robert-Tissot 2018) on the other hand. Since, as mentioned above, the turn-taking structure of WhatsApp-Chats also permits a discursive analysis, pragmatic factors are added to the analysis as well. This permits to single out the syntactic and discursive factors which trigger argument drop for each language under examination (namely the three national Romance languages of Switzerland: French, Italian

[1] These studies were conducted on the Swiss SMS corpus (cf. Stark/Ueberwasser/Ruef 2009-2015).

[2] The examples given in this paper are original WhatsApp or text messages and may contain orthographic errors, since they are informal performance data.

GG@G (Generative Grammar in Geneva) 11: 175-187, 2018

and Sursilvan as for the Romansh varieties) and to reveal and explain language specific or general regularities. The goal of this paper is to describe the annotation schemes I use to annotate the WhatsApp corpus and the theoretical background they are based on.

The paper is structured as follows: Section 2 presents two analyses of subject drop in written abbreviated registers mentioned above. Section 2.1. introduces the study on subject drop in English and French diaries by Haegeman, while section 2.2. introduces the research by Robert-Tissot on subject drop in French text messages. Section 3 briefly describes the data which is annotated for this research. Section 4 is dedicated to the different levels of annotation: The semantic level, which also comprises pragmatic factors, is presented in section 4.1., a first syntactic level (categories) is described in section 4.2., a second one (functions) is introduced in section 4.3. Section 5 summarizes the main ideas of this paper.

2. DIVERGING THEORIES ON ARGUMENT OMISSION IN NON-STANDARD WRITING

Several studies have examined the omission of subjects and objects in non-standard writing. With regard to abbreviated (mobile) registers, two diverging analyses have been proposed for subject drop: the truncation analysis by Liliane Haegeman (cf. section 2.1.) and the *topic* drop analysis by Aurélia Robert-Tissot (cf. section 2.2.).

2.1. Subject omission in English and French diaries

Haegeman's research examines subject drop in French and English diaries and is based on the analysis of null subjects in early child language production by Rizzi (1994, 2006). Her key observation is that omitted subjects trigger agreement on the inflected verb and are thus syntactically active. According to Haegeman, subject omission in abbreviated written registers does not correspond to any of the three classical types of argument drop (*pro* drop, *topic* drop and *discourse* drop, cf. Sigurðsson 2011), because it is subject to several distributional restrictions: it is restricted to main clauses, not possible in subject-auxiliary inversions, incompatible with fronted *wh*-elements or with fronted arguments, but compatible with preposed adjuncts (2). Haegeman stresses that her findings are similar to those found by Rizzi for early null subjects (Haegeman 1997:245).

(2) This morning Ø woke to get a letter in the mail […]
(cf. Haegeman 2017:232, edited)

Haegeman argues further that null subjects in diaries cannot be instances of *topic* drop, since *topic* drop also affects objects, which are not systematically omitted in her corpus. Furthermore, expletive subjects, which cannot be topical, are often dropped (cf. Haegeman 2017). Throughout her many years of research, Haegeman designates this phenomenon using several different terms: Starting with "null subject in diary context" (NSD) in 1997, she refers to it as "adult null subject" (ANS) in 2000 and as "diary subject omission" (DSO) in 2013. In her latest article on this topic (i.e. 2017), she names the phenomenon "Written Subject Omission" (WSO). Haegeman states that null subjects in diaries are a root phenomenon and models her findings within the framework of the phase-based theory of Generative Grammar (cf. Rizzi 2006), assuming a split CP as well as a SubjP (going back to Rizzi 1997 and Cardinaletti 1997 and 2004), the canonical subject position. She proposes that, for sentences with null subjects, the derivation ends at SubjP, which then becomes the root phase. Accordingly, the subject, due to its position in SpecSubjP, cannot be spelt out. To account for zero subjects with fronted adjuncts as in (2), she argues that the subject (*I*) is situated in the specifier of SubjP, whereas the adjunct (*this morning*) is adjoined to TP, thus occupying a lower position (cf. Haegeman 1997, 2013). The derivation is illustrated in (3), with the larger square brackets representing the end of the spell-out domain.

(3) [$_{SubjP}$ ~~I~~$_i$ [[$_{TP}$ this morning [$_{TP}$ t$_i$ woke to get a letter in the mail]]]
(cf. Haegeman 2017:242)

This is a promising approach for null subjects in diaries, but there are some caveats concerning this analysis. Above all, the analysis does not account for null objects. Nevertheless, several categories can be identified on the basis of this analysis for data annotation in order to test similarities between our corpus and WSO according to distributional restrictions: main and subordinate clauses, subject-auxiliary inversions as well as *wh*-elements, arguments and adjuncts in a subject-preceding position.

2.2. Subject omission in French text messages

Robert-Tissot (2018) investigates subject omission in French text messages. Her study revealed that the distribution of dropped subjects is very similar to that found in diaries by Haegeman, but it also manifests some important differences. Contrary to Haegeman, Robert-Tissot finds instances of dropped objects in her corpus (4). Moreover, she identifies dropped clitic subjects after fronted subject arguments (5) and even one instance of a dropped subject in an embedded clause (6) (cf. Robert-Tissot 2018:283 ff.).

(4) Alors c'est bon?t'as trouvé Ø?.
So it is good?you have found?
'So is it good? Did you find [it]?'

(5) Bref ben moi$_i$ mtn t$_i$ me$_k$ suis t$_i$ t$_k$ organisé différement […].
In short well me now myself am organized differently […].
'Actually me [I] organized myself differently now […].'

(6) Moi$_i$ j te cache pas qu t$_i$ t$_i$ suis pas mott à faire grd chose!
Me I you hide not that am not motivated to do big thing.
'Me, I won't hide from you that [I] am not motivated to do much.'

Example (5) shows a dropped clitic subject (*je*, 'I') after a fronted argument (the fronted strong subject pronoun *moi,* 'me'). *Moi* is topical and analyzed as a TopP, located higher in the left periphery than subjects (cf. Rizzi 1997), which rules out the truncation theory (cf. Stark/Robert-Tissot 2017). As for (6), she assumes that the null element is the strong subject pronoun *moi* at the beginning of the main clause, extracted from the subordinate clause[3] and interpreted as a contrastive topic (cf. Robert-Tissot 2018:315). This recalls long distance extraction of Foci in Hungarian (cf. i.e. Puskás 2000, Jánosi 2014). These drawbacks to the truncation analysis as well as the fact that Robert-Tissot also found instances of dropped objects in her corpus (cf. example (4)) lead her to argue for omitted referential subjects (and potentially also objects) as instances of *familiar topic* drop. A *familiar* topic is a given or accessible constituent which is typically distressed and realized in a pronominal form (cf. Frascarelli/Hinterhölzl 2007). Robert-Tissot argues that topics, of this kind at least, can be moved to the left periphery, leaving a trace in-situ. Then, *familiar* topics, being typically realized as clitic pronouns in French, cannot be spelled-out in the relevant position, due to the fact that they are separated from the verb, i.e. the host they would have to lean on (cf. Robert-Tissot 2018:301).

[3] For French, subject extraction from subordinate clauses is allowed under certain conditions (cf. Rizzi/Shlonsky 2007). I refer to Robert-Tissot (2018) for further details on her assumption.

Robert-Tissot thus provides a sound analysis for null subjects in French text messages that also permits an implementation on null objects. The fact that she finds similar distributional restrictions as those found in diaries by Haegeman underlines the necessity of the emerging categories for the annotation as mentioned in 2.1., for contrasting the distribution of null arguments in WhatsApp messages to text messages. Furthermore, a more detailed annotation of pronouns as clitic, strong and null allows for testing the *familiar topic* drop hypothesis. Yet, there is a problem with her analysis: Her proposal is not applicable for English, since English does not have a low TopP (cf. Bianchi/Frascarelli 2010, Haegeman p.c.).[4]

3. DATA

The data underlying my research stem from the Swiss WhatsApp corpus, which is still being worked on (Stark/Ueberwasser/Göhring 2014-), and were collected in Switzerland in June and July 2014. The population was invited via a media call in different newspapers and over the radio to contribute a copy of their actual WhatsApp chats to an internal link and e-mail address. In a second step, the contributors received access to a tool with which they could give their consent for their messages to be used anonymously for research. If they consented, they then where asked to fill out a questionnaire on demographic information.

The entire corpus contains more than one million messages, of which 763,650 messages written by 945 participants with consent (5,543,692 tokens). The four national languages of Switzerland are represented as follow (in decreasing order): Swiss German (506,984 messages), French (197,255 messages), standard German (81,456 messages), Italian (42,559 messages) and Romansh (29,094 messages). For more detailed information on the corpus, see Ueberwasser/Stark (2017). For the present study, a smaller sub-corpus consisting of 4000 French messages, 1000 Italian messages and 1000 Sursilvan messages is annotated on different levels, which are presented in the following section.

4. ANNOTATING ARGUMENT DROP IN WHATSAPP MESSAGES

For the annotation, only messages with a finite verb, as in (1), and only sentences where null arguments can be reconstructed are considered, so-called telegraphic style messages, like the one in (7), have been discarded.

(7) Demain aller au ciné?
Tomorrow go to cinema?
'Tomorrow go to the movies?'

All arguments of the finite verb and the finite verb itself are considered. Verbal complements (direct objects, indirect objects, prepositional objects and adverbial complements) are identified via the respective argument structure and subcategorization frame of the verb. In cases of highly polysemous verbs like *passer* (e.g. 'cross', 'pass by', 'get through'), the intended meaning of the verb in the given context and its respective argument structure is taken as the basis of the annotation. Adjuncts are annotated only in those cases in which they are situated at the beginning of the sentence. Standard omissions like in imperatives as well as parentheses are ignored. Dropped arguments are, where possible and given by context, analyzed as pronominals (cf. Sigurðsson 2011:289). Thus in (1), a missing direct object clitic *le* ('it') is assumed to be present implicitly, constituting the direct object of the transitive verb *finir* ('end something') and coreferential with the preceding *ppx* ('power

[4] Thanks to Liliane Haegeman for discussing this point with me.

point presentation'). Alternatively, a nonspecific referent such as 'the work', 'what I have to do' etc. could be assumed to be the direct object of *finir*, which I did not consider, given the possibility of analyzing it as a pronominal. Finally, *finir* could also be considered to be intransitive here ('to have arrived at the end of something'), which would render the subordinate sentence *si j'ai fini* ('if I have arrived at the end') complete. In this and similar cases, I decided to follow the *Grand Robert de la Langue francaise* (http://gr.bvdep.com/) by checking the possible argument structures of the verb under investigation. In a second step, I interpreted the intended meaning in the given context.

All verbal arguments in the messages of my subcorpus are annotated manually with the annotation program MMAX2 (Müller/Strube 2006). The annotation schemes are based on the syntactic constraints identified in the work of Haegeman and Robert-Tissot (cf. section 2) as well as on pragmatic factors geared towards identifying the 'accessibility' status of the respective discourse referents (cf. Ariel 1988) and therefore making use of the interactional character of the chats. The individual annotation categories and factors are described in detail below.

To facilitate the annotation process, the schemes are constructed in such a way that they can be used for cross-linguistic comparative analyses[5]. The annotation is performed on three independent levels: on a semantic level (cf. section 4.1.), where semantic and pragmatic properties of each argument are labelled, as well as on two syntactic levels, one for the categories (cf. section 4.2.) and another for the functions (cf. section 4.3.). Moreover, MMAX2 provides the user with a tool to indicate relations or dependencies between two markables, which is used on the semantic level and the function level and is explained further in the relevant sections. Within the program, the base data (the messages) can be modified for inserting zero elements, which can then be annotated on each level like the other elements.

4.1. The semantic level

This level comprises two sub-levels that are annotated individually for each argument: the first one for information structural properties and the second one for the referential category. As shown in figure 1, the information structure level marks whether the argument is newly introduced or given. The second one is divided further according to the referential distance to the last explicit coreferential element of the argument (in the same sentence, in the same message, in the preceding message or in the further preceding discourse; cf. the four categories of Ariel 1988: same sentence, previous sentence, same paragraph, across paragraph). This allows for a more fine-grained grasp of (topic) accessibility and the correlating encoding (cf. Givón 1983, Ariel 1988, Prince 2006). Furthermore, it allows for testing of Robert-Tissot's hypothesis that omitted arguments are *familiar topics* (i.e. have an antecedent in the preceding discourse and are highly accessible). According to Givón, the closer the antecedent, the higher the accessibility of the discourse referent and the lesser the material with which it is verbalized. Thus, zero anaphors should be found when referents are highly accessible, unstressed pronouns whenever they are less accessible, stressed pronouns when they are barely accessible and full NP's (cf. Givón 1983:17f.) when there is no referent at all, i.e. the argument is new. On this level, a relation type attribute can be added for each argument, relating it to its first explicit coreferential element. This subsequently permits to study topic chains or anaphoric chains in the sense of Givón. If the data show the patterns described by Givón or Ariel, argument drop in WhatsApp messages is hence not arbitrary.

[5] Therefore, in each figure, language-specific values are pointed out in italics.

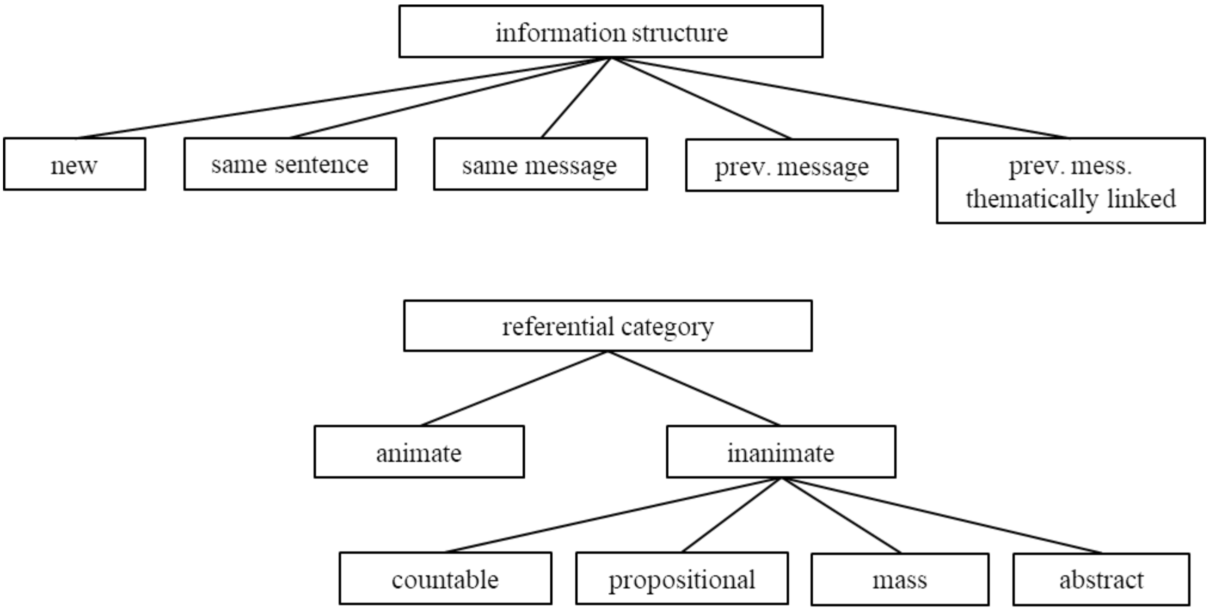

Fig. 1: Values on the semantic level

As for the referential category, each argument is labelled as to its animacy. If the argument is inanimate, it is further annotated as countable, propositional, mass or abstract. Since countable does not exclude abstract, the arguments are in addition annotated as to their ontological category intended in the given context (i.e. not interpreting metaphors). Specifically, this means that *un amour* ('a love') would be annotated as countable in a context where it is mentioned as exactly one love, designating a concrete relationship for example, but as abstract in contexts where it denotes an idea or a concept. Mass, on the other hand, is assigned to all arguments that describe uncountable entities, like *sel* ('salt'). Nouns like *sucre* ('sugar') that can have both mass and countable readings are again annotated according to the intended meaning. The value 'propositional' is assigned to arguments like French *ceci* ('this') when referring to a preceding matter.

The referential categories are important for testing the relation between animacy and argument omission, that have been shown to be linked for Brazilian Portuguese: In this language, null objects are usually inanimate (cf. Cyrino 1997, Cyrino/Matos 2016). The results of the corpus search for the referential category permit to check if the features of a null argument (i.e. null object) are similar to those found in Brazilian Portuguese. Furthermore, Fernández-Ordóñez (2012:97) found a distinction between mass and count for some Ibero-Romance pronoun paradigms; this, too, could be identified with the help of this scheme, as well as a possible difference between propositional, abstract and mass arguments.

As mentioned before, MMAX2 provides a tool to signal relations or dependencies between two markables. To be able to query for chains of coreferential expressions, such markable_pointer-type relations are set on this level between an argument (the source markable) and its first coreferent expression (the target markables), if applicable (a newly introduced argument does not point to any preceding coreferential element because it does not have one). Example (1) shows such a relation by the indices: in the program, the zero element would be connected to the coreferential expression *un ppx* ('a power point presentation') by a line.

4.2. Syntactic level – categories

In a second step, the syntactic categories of each element have been classified. This level is composed of the following attributes: pronoun, proper noun, determiner phrase (DP), noun phrase (NP), quantifier phrase (QP), complementizer phrase (CP), verb phrase (VP), tense phrase (TP), adverb and preposition phrase (AdvP/PP) as well as verb. Two more options which are important for this study have been added: one for a zero element which is not pronominal, as well as one for elements that are replaced by an emoji. This layer ties in with the semantic level, since it encodes the amount of phonological material (i.e. null, clitic, stressed pronoun, full phrase) of each argument, which is, according to Givón, related to accessibility (cf. section 4.1.).

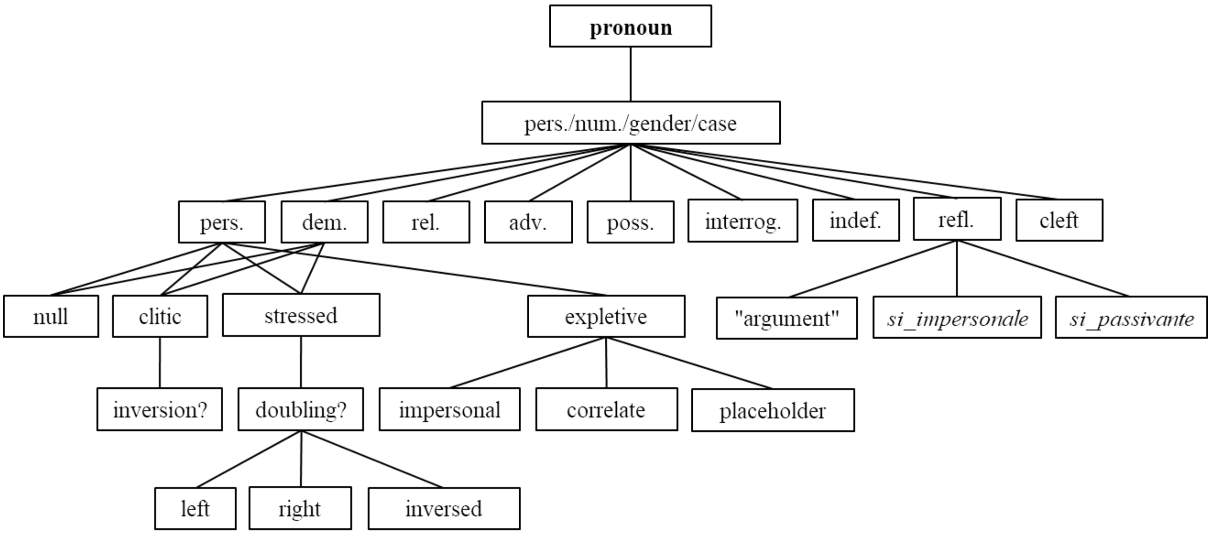

Fig. 2: Non-verbal constituents on the syntactic level – categories: pronouns

Figure 2 shows all the values that can be assigned to a pronoun (note that it does not have to be interpreted as hierarchical). The pronouns are classified into personal, demonstrative, relative, adverbial, possessive, interrogative, indefinite, reflexive and expletive found in clefts (including pseudo-cleft). In order to simplify the annotation process, cleft expletives are grouped among pronouns, and reflexive pronouns received their own label (though they are actually a sub-group of personal pronouns). Where necessary, person, number, gender and case are annotated. Person, number and gender are considered for personal, relative, possessive and reflexive pronouns, case is considered for personal, relative and reflexive pronouns. The third person singular pronoun *on* ('we') in French is annotated separately in order to distinguish it from the first person plural pronoun *nous* ('we'). Personal and demonstrative pronouns can be null, clitic (*je*, 'I', *ce* 'this') and stressed (*moi*, 'me', *ça*, 'that'). Clitic pronouns are further annotated as to whether they are realized in subject verb inversion contexts or not (as in questions like *Aimes-tu le chocolat*, 'Do you like chocolate', or in exclamations like *Quelle veinarde suis-je*, 'Such a lucky duck am I') whereas stressed pronouns are further labelled according to whether they are right- (*Tu as fait quoi toi?*, 'You what did you do?'), left-dislocated (*Moi je t'embrasse*, 'Me I hug you'), or inversely doubled (*Lui est-il venu?*, 'He did he come?'). Personal pronouns have the additional possible feature 'expletives'. Expletive pronouns are further divided into impersonal (comprising the subjects of impersonal verbs like *Il y a*, 'There is/are' and weather verbs like *Il fait froid*, 'It is cold'), correlate (*C'est bien que…*, 'It is good that…') and placeholder (*Il a été commandé trois verres de vin*, 'It was ordered three glasses of wine'). The subject of an idiomatized expression (e.g. *ça depend,* 'it depends') is labelled as impersonal. This distinction is made on the basis of phoricity: While correlates are phoric, impersonals and placeholders are not. Placeholders, however, are not arguments of the verb. I therefore assume that the different kinds of expletive pronouns cannot be dropped in a similar way. A reflexive pronoun is annotated by default as 'argument', but can also be assigned the categories *passivante* (*In questo ristorante* *si mangia bene*, 'In this restaurant one eats well') and *impersonale* (*In Italia si parlano molti dialetti*, 'In Italy many dialects are spoken'). Consequently, in searching personal pronouns in the corpus, in order to get all instances, one must query for personal pronouns and reflexive pronouns, since the two are annotated as distinct categories.

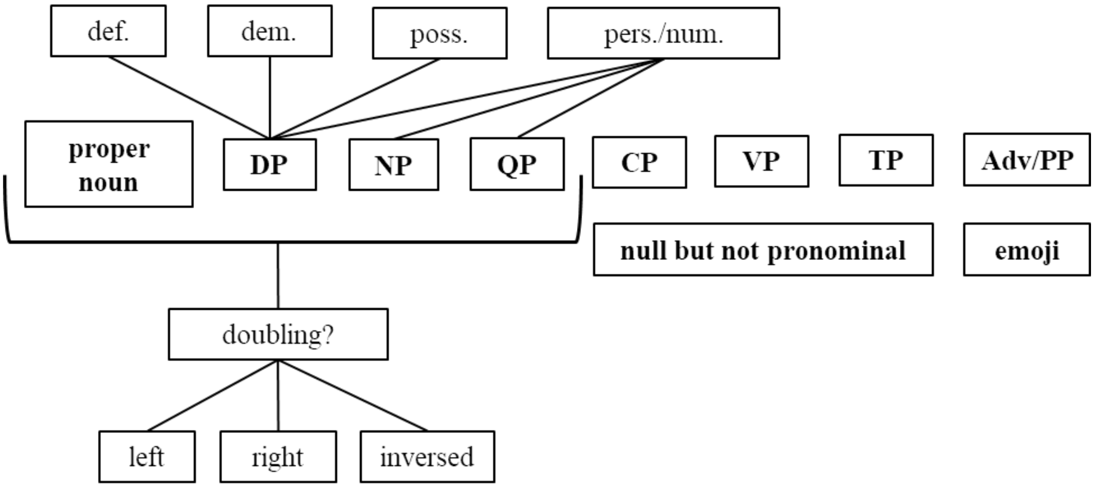

Fig. 3: Non-verbal constituents on the syntactic level – categories

Non-pronominal arguments, shown in figure 3, are not annotated in as much detail as the pronouns. The category 'proper noun' comprises anthroponyms, toponyms and other proper names such as movie titles. A DP is annotated according to its determiner as definite (*le chien*, 'the dog'), demonstrative (*ce chien*, 'this dog') or possessive (*mon chien*, 'my dog'). QPs include phrases with quantifiers like indefinite articles (*un chien*, 'a dog'), 'partitive articles' (*du chien*, 'dog'), numerals (*cinq chiens*, 'five dogs'), universal quantifiers (*chaque chien*, 'every dog'), negative quantifiers (*aucun chien*, 'no dog') and mid-scalar quantifiers (*quelques chiens*, 'some dogs'). A doubling feature is available for proper nouns (*Anne, elle est partie*, 'Anne, she left'), DP's (*Ce livre, je l'ai lu*, 'This book, I read it') and NP's (*Blade Runner, je l'ai bien aimé*, 'Blade Runner, I liked it a lot'). DP's, NP's and QP's are also annotated as to their number. For statistics, one has to pay attention that when searching the corpus for the third person, the relevant personal pronouns as well as the proper nouns, DP's, NP's and QP's have to be queried. Two more categories are available on this level: The value 'null but not pronominal' can be assigned to zero nominal elements where a reconstruction of a pronoun can definitely be excluded, as in *je pars manger Ø,* 'I go to eat'. The value 'emoji', indispensable for this study, since nominal constituents can be realized as emojis in WhatsApp messages because of the equivalent and new feature on smartphones in general and in the WhatsApp application, can be assigned to every constituent which is replaced by an emoji, (*Il est où le* 🐶? 'Where is it, the dog?').

The verbal constituents are labelled as to their verbal morphology and some semantic classes (see fig. 4). For each verb, mode (indicative, subjunctive, conditional and imperative) and tense (synthetic: present, imperfect, *passé simple*, future; analytic: perfect, pluperfect, future compound, future perfect) have been defined as it appears in the message. For analytic tenses, the possible auxiliaries *avoir_avere*, 'have' and *être_essere*, 'be', as well as *aller_andare*, 'go' for future compound can be chosen. Furthermore, a null value for dropped auxiliaries as in *ieri Ø Ø andato al cinema*, 'yesterday [I] [have] gone to the movies' is applicable. Concerning verbal morphology, agreement features have not been annotated, since subject-verb-agreement is not the focus of this study. The relevant information can be retrieved from the labelling of the subject. In case the inflection of the verb does not correspond to the subject, I assume that it is a case of misspelling.

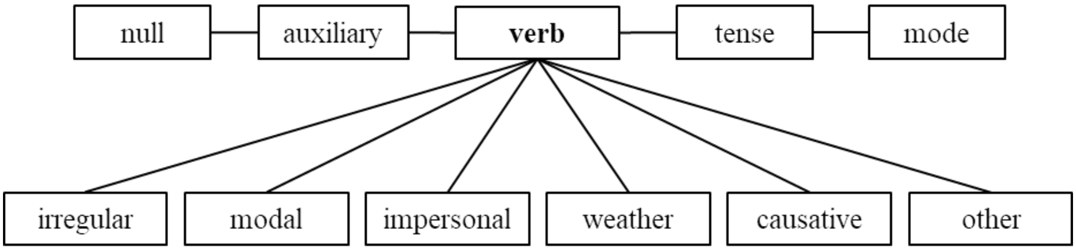

Fig. 4: Verbal constituents on the syntactic level – categories

Verbs are then classified into several semantically and morphologically defined groups (see fig. 4), the latter comprising morphologically irregular verbs (strong allomorphy suppletive paradigms) such as *avoir_avere*, 'have', *être_essere*, 'be' and *aller_andare*, 'go'. The semantic classes consist of modal verbs, impersonal verbs, weather verbs and causative verbs. If a verb does not fit into one of these groups, it is labelled as 'other' and its infinitive is written down in a freetext string. The attribute 'modal' may have the values *vouloir_volere*, 'will', *pouvoir_potere*, 'can' and *devoir_dovere* 'must'. The attribute 'impersonal' can have the following values: *falloir_bisognare*, 'need', *y avoir_esserci*, 'there is', *s'agir de_trattarsi di*, 'be about', *suffire_bastare*, 'be enough' and in their impersonal use *sembler_sembrare*, 'seem', *paraître_parare*, 'appear', *convenir_convenire*, 'suit'. Weather verbs are *pleuvoir_piovere*, 'rain', *neiger_nevicare*, 'snow', *grêler_grandinare*, 'hail', *tonner_tuonare*, 'thunder' as well as *geler_congelare*, 'freeze'. Causative verbs are *faire_fare*, 'make' and *laisser_lasciare*, 'let'. Each list being not exhaustive, the value 'other' can be assigned to a new lexical item not yet comprised in the respective list. Its infinitival form is entered manually in a freetext string.

To sum up, this level is used to obtain in-depth information on the syntactic categories of each element in the WhatsApp message, as well as on the influence of different types of verbs. These categories are annotated in order to identify morphological and semantic features (such as expletives) that may correlate with omission and to provide comprehensive illustrations of the diverse properties that might influence argument drop in general.

4.3. Syntactic level – functions

On the second syntactic level, the elements can be annotated according to their function in the clause, including the elements in the left periphery. The level of syntactic functions (see fig. 5) is composed of the following attributes: subject, predicate, direct object, indirect object, prepositional object, adverbial complement and predicative.

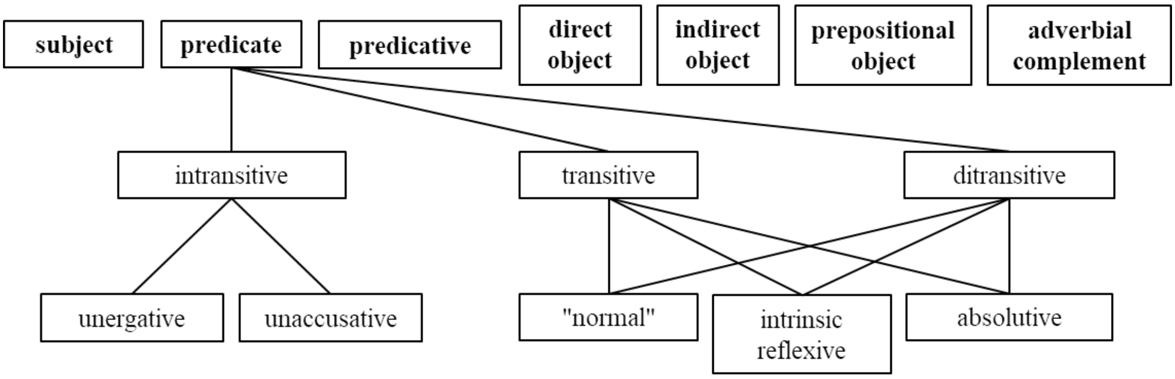

Fig. 5: Values on the syntactic level – functions

The attribute 'subject' is assigned to the element in subject position. No further details are annotated on this attribute, since they can be recovered by looking at elements assigned nominative case on the syntactic level – categories. Predicates are classified according to their argument structure and the meaning intended in the given context. The values 'intransitive', 'transitive' and 'ditransitive' can be assigned, depending on the subcategorization of the predicate. If a verb is used intransitively, it can be distinguished further into 'unergative' (verbs like *dormir,* 'sleep', *pleurer,* 'cry', *rire,* 'laugh') and 'unaccusative' (verbs like *arriver,* 'arrive', *partir,* 'leave', *mourir,* 'die', as well as raising verbs). If the verb under investigation is transitive (copular or functional like *venir de* plus infinitive ('have just done'), modal verbs and causative verbs) or ditransitive, it can be marked as to whether its realization in the sentence is normal, whether it is intrinsically reflexive as in *se réjouir,* 'be happy' or *pentirsi,* 'regret' or whether it is absolutive, meaning with no or only one complement instead of one or two, as shown in examples (8), (9) and (10)[6].

(8) Je pars manger Ø.
I leave eat.
'I am heading out to eat [something].'

(9) Je sais pas Ø.
I know not.
'I don't know [it].'

(10) Tu me Ø redis demain.
You me tell again tomorrow.
'You tell me [it] again tomorrow.'

As for the verbal complements, the three functions of direct object, indirect object and prepositional object as well as the ones of adverbial complement and predicative complement can be marked on each argument.

Note that one and the same verb can have different argument structures. For example, the verb *penser*, 'think', can be used either intransitively (as in *sa manière de penser,* 'his/her way of thinking'), transitively with a direct object (*je te dis ce que je pense,* 'I tell you what I think') or transitively with a prepositional object (*je pense fort à toi,* 'I think about you all the time'). For the annotation, one must bear in mind the possible argument structures of each verb and carefully look at the use employed in the message.

In order to conduct more efficient queries after annotation, markable_pointer-type relations are set on this level between the verb (the source markable) and its arguments (the target markables). For example, in (10), *redis* ('tell again') points to *tu* ('you') as its subject, to *me* ('me') as its indirect object and to the null element as its direct objet. This step simplifies the subsequent querying process, because it allows searching for combined elements. If such a relation is not set and one wished to see all transitive verbs with a zero direct object as well as all these zero direct objects, the results would show two independent lists with transitive verbs with a zero direct object as well as all zero direct objects, but would not signal which object belongs to which verb. A pointer relation indicates this very dependency and can be queried directly, resulting in a list with all matched verb-direct object pairs. Moreover, this relation allows queries for whether the argument follows or precedes the verb (and thus allows to test for one of the distributional restrictions found by Haegeman, mentioned in section 2.1.).

[6] See also the absolutive use of *finir* ('finish') in example (1).

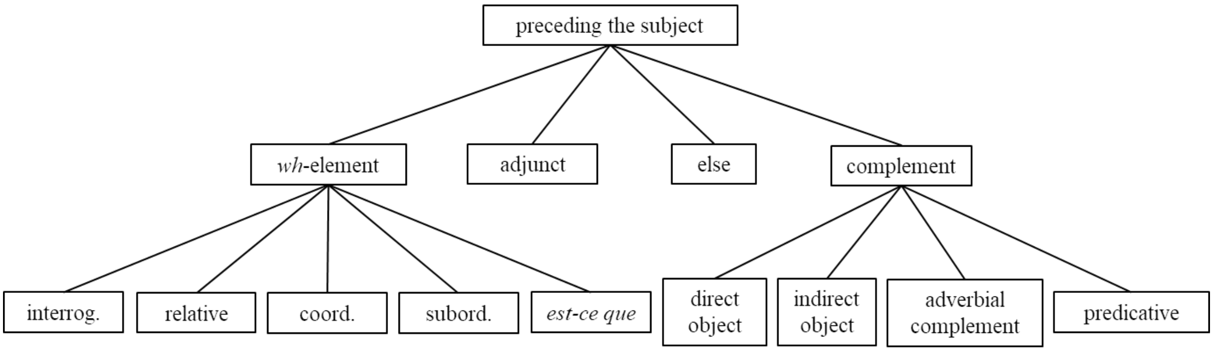

Fig. 6: Elements in the left periphery

Finally, figure 6 shows the elements in the left periphery potentially preceding zero subjects, thereby permitting to test for the validity of the truncation hypothesis and for the distributional restrictions that Haegeman found for null subjects in English and French diaries (cf. section 2.1. above). I repeat these restrictions as listed in her recent paper (cf. Haegeman 2017): Null subjects in English and French diaries are restricted to root clauses and unavailable with fronted arguments, *wh*-elements and with subject auxiliary inversions. However, they are compatible with fronted adjuncts. As shown in figure 6, all these elements in the left periphery can be marked: *wh*-elements, including relative pronouns, interrogative pronouns and French *est-ce que* as well as subordinate (complementizers) and coordinate conjunctions (for the sake of simplicity); preceding adjuncts and, most importantly, preceding arguments (rather complements, including direct and indirect objects, adverbial complements, predicatives and prepositional objects). The option 'other' is chosen for certain discourse elements like French *alors,* 'so'*, donc,* 'so', *en fait,* 'actually'. Interjections and onomatopoeia are not considered. If the preceding element is a complement, it also receives its properties on the semantic level, including a pointer relation to its first coreferential expression, as explained in 4.1.

These markables, once set, also serve as target markables for markable_pointer-type relations, with the sentence subject as the source markable. As explained for the coreferential expressions and syntactic functions, a pointer relation between the subject and a preceding element allows for querying these two categories in combination and not separately.

All in all, the syntactic functions are annotated to gain a detailed understanding of the interplay of argument omission and transitivity as well as distributional restrictions of (null) arguments, which would indicate that the omissions follow syntactic regularities and cannot be explained by purely extralinguistic or D-linking factors.

5. CONCLUSION

This study aims to provide a systematic empirical investigation on (null) arguments in (Romance) WhatsApp messages, a relatively new form of written mobile communication. In order to find out whether argument drop in WhatsApp messages is technically or syntactically motivated (maybe as a register-specific feature), the results of this study will be compared to those that emerged from the study on the Swiss SMS corpus (cf. Robert-Tissot 2018). Moreover, two analyses of subject drop in written registers will be tested: the truncation analysis by Haegeman (2013, among others) and the *topic drop*-analysis by Robert-Tissot (2018). For this purpose, the relevant descriptive categories emerging from their approaches (main vs. subordinate clause, *wh*-elements, fronted arguments and adjuncts, subject-auxiliary inversion, strong and weak pronouns) are annotated on the data. These categories have been added to detailed annotation schemes that also include other grammatically relevant properties, like semantic categories, syntactic constituents and functions, different kinds of verbs and pronouns etc. One difference between the different communication forms in question (diaries, text messages and WhatsApp messages) should not be underestimated: WhatsApp messages are more interactional than the other two, and in the Swiss WhatsApp corpus, the surrounding context of the elements under scrutiny is given, which is also reflected in the annotation scheme: For each argument, the distance to its first explicit coreferential element is annotated. This allows for testing the topicality-continuity categories of Givón (1983) or the accessibility scale by Ariel (1988).

By proceeding in such a way, the syntactic, pragmatic and functional distribution of each argument can be described. This provides a much broader database for the subsequent cross-linguistic quantitative and qualitative analyses, hopefully yielding the discovery of the regularities underlying argument drop in mobile electronic written communication.

REFERENCES

Ariel, M. (1988) "Referring and accessibility", *Journal of Linguistics* 24:1, 65-87.

Bianchi, V. & M. Frascarelli (2010) "Is Topic a Root Phenomenon?", *Iberia* 2:1, 43-88.

Cardinaletti, A. (1997) "Subject and clause structure", in L. Haegeman (ed.) *The New Comparative Syntax*, Longman, London/New York, 33-63.

Cardinaletti, A. (2004) "Toward a Cartography of Subject Positions", in L. Rizzi (ed.) *The Structure of CP and IP*, Oxford University Press, New York, 115-165.

Fairon, C. & J.R. Klein, S. Paumier (2006) *Le langage SMS*, Presses universitaires de Louvain, Louvain-la-Neuve.

Fernández-Ordóñez, I. (2012) "Dialect areas and linguistic change", in G. De Vogelaer & G. Seiler (eds.) *The Dialect Laboratory*, John Benjamins, Amsterdam/Philadelphia, 73-106.

Frascarelli, M. & R. Hinterhölzl (2007) "Types of topics in German and Italian", in K. Schwabe & S. Winkler (eds.) *On Information Structure, Meaning and Form*, John Benjamins, Amsterdam/Philadelphia, 87-116.

Frick, K. (2017) *Elliptische Strukturen in SMS. Eine korpusbasierte Untersuchung des Schweizerdeutschen*, de Gruyter, Berlin/Boston.

Givón, T. (1983) "Topic continuity in discourse. An introduction", in T. Givón (ed.) *Topic continuity in discourse. A quantitative cross-language study*, John Benjamins, Amsterdam/Philadelphia, 1-41.

Haegeman, L. (1997) "Register variation, truncation, and subject omission in English and in French", *English Language and Linguistics* 1, 233-270.

Haegeman, L. (2000) "Adult null subjects in non pro-drop languages", in M.-A. Friedemann & L. Rizzi (eds.) *The Acquisition of Syntax: Studies in Comparative Developmental Linguistics*, Longman, London/New York, 129-169.

Haegeman, L. (2013) "The syntax of registers: Diary subject omission and the privilege of the root", *Lingua* 130, 88-110.

Haegeman L. (2017) "Unspeakable sentences: Subject omission in written registers: a cartographic analysis", *Linguistic Variation* 17:2, 229-250.

Ihsane, T. (1998) *The syntax of diaries: grammar and register variation*, University of Geneva. Ms.

Janda, R.D. (1985) "Note-taking English as a simplified register", *Discourse Processes* 8, 437-454.

Jánosi, A. (2014) *Long Split Focus Constructions in Hungarian with a View on Speaker Variation*, Catholic University of Leuven. Ms.

Massam, D. & Y. Roberge (1989) "Recipe context null objects", *Linguistic Inquiry* 20, 134-139.

Müller, C. & M. Strube (2006) "Multi-Level Annotation of Linguistic Data with MMAX2" in S. Braun & K. Kohn, J. Mukherjee (eds.) *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods,* Peter Lang, Frankfurt, 197-214.

Panckhurst, R. (2009) "Short Message Service (SMS). Typologie et problématiques futures", in T. Arnavielle (ed.) *Polyphonies, pour Michelle Lanvin*, Université Paul Valéry Montpellier, Montpellier 3, Montpellier, 33-52.

Prince, E.F. (2006) "Impersonal Pronouns in French and Yiddish" in B.J. Birner & G. Ward (eds.) *Drawing the boundaries of meaning*, John Benjamins, Amsterdam/Philadelphia, 295-315.

Puskás, G. (2000) *Word Order in Hungarian: The Syntax of A'-positions*, John Benjamins, Amsterdam/Philadelphia.

Rizzi, L. (1994) "Early null subjects and root null subjects", in T. Hoekstra & B.D. Schwartz (eds.) *Language Acquisition Studies in Generative Grammar*, John Benjamins, Amsterdam/Philadelphia, 151-176.

Rizzi, L. (2006) "Grammatically-based target-inconsistencies in child language", in K.U. Deen et al. (eds.) *The Proceedings of the Inaugural Conference on Generative Approaches to Language Acquisition - North America (GALANA)*, MIT Press, Cambridge, Mass., 19-49.

Rizzi, L. & U. Shlonsky (2007) "Strategies of Subject Extraction", in H.-M. Gärtner & U. Sauerland (eds.) *Interfaces + recursion = language? Chomsky's minimalism and the view from syntax-semantics*, Mouton de Gruyter, Berlin,115-160.

Robert-Tissot, A. (2018) *Grammaire du SMS*, Presses universitaires de Vincennes, Vincennes.

Sigurðsson, H. (2011) "Conditions on argument drop", *Linguistic Inquiry* 42, 267-304.

Stark, E. (2016-2018) *SNSF project "What's up, Switzerland?"* (Sinergia: CRSII1_160714), University of Zurich. www.whatsup-switzerland.ch.

Stark, E. & P. Meier (2018) "Argument drop in Swiss WhatsApp messages – A pilot study on French and (Swiss) German", *Zeitschrift für französische Sprache und Literatur* 127:3, 224-252.

Stark, E. & A. Robert-Tissot (2017) "Subject drop in French text messages", *Linguistic Variation* 17:2, 251-271.

Stark, E. & S. Ueberwasser, B. Ruef (2009-2015) *Swiss SMS Corpus*, University of Zurich. www.sms4science.ch.

Stark, E. & S. Ueberwasser, A. Göhring (2014-) *Corpus "What's up, Switzerland?"*, University of Zurich. www.whatsup-switzerland.ch.

Ueberwasser, S. & E. Stark (2017) "What's up, Switzerland? A corpus-based research project in a multilingual country", *Linguistik Online* 84:5.

Weir, A. (2017) "Object drop and article drop in reduced written registers", *Linguistic Variation* 17:2, 157-185.