Annotation guidelines ANCORsyn

Eric Engel

10/18/2022

Contents

1 Introduction

These guidelines present the annotation procedure for ANCORSYN, an adapted and extended version of part of the ANCOR_Centre corpus.

ANCOR_Centre is a corpus of spoken French comprising 488000 words, made available under a Creative Commons licence (CC-BY-SA for the subcorpora OTG, Accueil_UBS, and ESLO-CO2; and CC-BY-SA-NC for the ESLO_ANCOR subcorpus) (Muzerelle et al. 2013, 2014). The version used in this work is v1.1, which has been retrieved from the Ortolang repository on https://hdl.handle.net/11403/ortolang-000903/v1. The data is also available from the corpus website at https://www.info.univ-tours.fr/~antoine/parole_publique/ANCOR_Centre/index.html, from where it can be downloaded without restrictions. It contains annotations of referential relations, along with some grammatical and semantic features of referring expressions (gender, number, part-of-speech of the head, inclusion in a prepositional phrase, named entity type, definiteness, and newness). Furthermore, the data is segmented into thematic sections and turns.

For the purpose of this work, four files (004_-1.xml, 010_C-1.xml, 013_C-1.xml, and 026_C-1.xml) have been selected from the ESLO_ANCOR subcorpus, which contains semi-guided sociolinguistic interviews. These files were converted from the project's XML format (annotation_integree) to the MMAX2 format (https://github.com/ottiram/MMAX2, cf. Müller & Strube 2006) for further analysis. The MMAX2 files contain stand-off annotations on five layers (`section`, `turn`, `edu`, `reference`, `disfluency`), which are outlined below.

Compared to the original annotation, the main differences include an additional segmentation layer to Elementary Discourse Units, a markup of disfluent speech, and a revised annotation

of referring expressions and their attributes. The modified resource is made available online[1] under the terms of the same license as the original data (CC-BY-SA-NC)[2], as required by that license. This means that the corpus can be used and transformed for non-commercial purposes, as long as the original authors are cited and the modified work is redistributed under the same conditions (i.e., the same CC-BY-SA-NC license).

Note that the reference publications of ANCOR_Centre Muzerelle et al. (2013, 2014) must be cited when using the data.

[1] https://github.com/erengel/ancorsyn
[2] See https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en

2 Thematic sections (section)

2.1 Markables

The ANCOR_Centre corpus contains markup of thematic sections, where one section corresponds to one question from the sociolinguistic interview guidelines including follow-up questions by the interviewer and the interviewee's answers. This segmentation was taken over as-is.

2.2 Attributes

- N: Sequence number of the section within the file.
- ID_TOPIC: The ID of the topic of conversation, as identified by the interview guidelines and indicated in the ANCOR_Centre files.

3 Conversational turns (turn)

3.1 Markables

ANCOR_Centre also includes speaker turns, where one turn corresponds to an uninterrupted stream of speech by one speaker, although overlapping speech has been segmented into a separate turn and assigned to both speakers. The turn annotation was revised in cases of overlapping speech, such that (i) each token is assigned to exactly one speaker, and (ii) instead of constituting a separate turn, overlapping speech was integrated into preceding or following turns by the same speaker wherever possible.

3.2 Attributes

- N: Sequence number of the turn within the file.
- ID_SPEAKER: ID of the person speaking.

4 Elementary Discourse Units (edu)

4.1 Markables

The following guidelines are based on proposals by Carlson et al. (2003), Muller et al. (2012) and Stede et al. (2017). Elementary Discourse Units (EDUs) are segments below the level of turns that serve a communicative purpose. As with turns, every token must be assigned to one and only one EDU. One EDU corresponds to one illocution. EDUs thus trigger updates of the discourse model by adding a proposition, updating a previously added proposition, or guiding the further development by asking a question.

The most basic case of an EDU is a finite verb plus all of its arguments.

(1) (euh) je travaillais chez Simca
'I worked at Simca.' [004_-1: 19]

The assignment of tokens to EDUs is exhaustive, in that every token must be assigned to exactly one EDU. For this reason, EDUs are necessarily non-overlapping, although they can be discontinuous (see (7) below).

4.1.1 Argument clauses and adverbial clauses

Embedded clauses are treated differently depending on their syntactic status: Argument clauses, such as the object clauses in (2) and (3), complete the proposition introduced by the matrix verb, and are therefore integrated into the larger EDU. Adverbial clauses, however, are

segmented into separate EDUs (as illustrated by the pipe/vertical bar |), since they express their own complete proposition, as in (4) or (5).

(2) j'espère que ce sera ces vacances
'I hope it's going to be this summer break.' [013_C-1: 188]

(3) je pense pas qu'il y ait des matières inutiles
'I don't think there are useless school subjects.' [013_C-1: 264]

(4) oui (je m'en) je m'en rends compte | parce que je viens souvent à Orléans
'Yes, I'm aware of that | because I often come to Orléans' [013_C-1: 441–442]

(5) moi enfin j'avais dix ans | quand j'ai commencé à en faire
'I myself was ten | when I started it.' [013_C-1: 294–295]

Note that in the case of quoted speech, the first part of the quoted speech that qualifies as an EDU is integrated into the matrix EDU containing the verb of saying, see (6), even if the quotation continues after that. This follows the general rule that clausal arguments of verbs do not constitute separate EDUs.

(6) et un jour (euh) ce monsieur cuisinier (me) vient me voir | et puis me dit (euh) "tiens Madame (euh) X a téléphoné | en me demandant quel jour vous pourriez la recevoir | elle a absolument besoin de vous voir"
'And one day, this cook came to me | and said "Miss X called me | asking what day you would be able to see her | she absolutely has to see you".' [010_C-1: 359–362]

EDUs can be discontinuous in cases where an adverbial clause appears in a clause-internal position (so-called "center-embedding"), or at the beginning of an object clause. This is not a problem when annotating in MMAX2: For an utterance like (7), the annotator first creates a markable on the EDU layer for the span *je vous dirai que* and then adds the span *le travail s'organise très bien* to it, before creating a second markable for the span *quand on est chez une*

petite couturière. However, one should be careful when converting the data for use with other annotation tools, as not all of them support discontinuous markables.

(7) je vous dirai que ...| quand on est chez une petite couturière |... le travail s'organise très bien. 'I'm gonna tell you that | when you're working in a small sewing facility, | you can organize the work very well.' [010_C-1: 261–262]

4.1.2 Relative clauses

In the segmentation of EDUs, we distinguish restrictive relative clauses, which are used to further specify their nominal head, and non-restrictive relative clauses, which provide additional information about an otherwise independently referring expression. Relative clauses mostly modify noun phrases referring to generic or previously unidentified entities, adding more content to the referring expression in order to identify the intended referent, as in (8), or provide sufficient content to establish a mental representation, as in (9). They are therefore analyzed as part of the larger EDU.

(8) du reste (euh) le pont que vous avez là tout à côté s'appelait, oui, le pont du collège
'By the way, the bridge that you have there right next to it was called "pont du collège".' [010_C-1: 37]

(9) je peux pas me donner (ce ce) cette qualification-là mais enfin (euh) les gens qui connaissent mon travail
'I cannot give that label to myself, but the people who know my work.' [010_C-1: 189]

In contrast, non-restrictive relative clauses add information to a previously identified entity, as in (10), where the daughter has been talked about before. Since the relative clause in these cases is not needed to establish the intended reference, but adds an independent piece of information, non-restrictive relative clauses are analyzed as constituting their own EDU.

(10) d'ailleurs ma fille | qui est infirmière psychiatrique | vous dira qu'ils ont bien plus de dépressions nerveuses chez des femmes (inoccupées, enfin,) sans travail qu'avec des personnes travaillant
'Also, my daughter | who's a psychiatric nurse | will tell you that they have many more cases of clinical depression among unemployed women than among people who are working.' [026_C-1: 178–179]

4.1.3 Cleft sentences and dislocations

One exception to the segmentation of non-restrictive relative clauses are cleft sentences. In cleft sentences like (11), the relative clause *qui m'inquiétait le plus* is non-restrictive, because it does not provide information needed to establish the reference of *cette chose-là.* However, the matrix clause *c'était surtout cette chose-là* does not constitute an independent proposition, but the structure *c'est + X qui …* serves to structurally mark *X* as the focus of the utterance. For these reasons, we do not adopt a biclausal analysis of clefts, and analyze the whole cleft as one EDU.

(11) c'était surtout cette chose-là qui m'inquiétait le plus
'It was especially that thing that I worried about most.' [010_C-1: 437]

Dislocated elements, including "hanging topics", are also integrated into the adjacent EDU. Their status as elements that are peripheral to the core clause can be retrieved from the annotation of syntactic roles (cf. SYNROLE in 5.2.1).

(12) et moi je suis restée à Orléans
'And I stayed in Orleans.' [026_C-1: 22]

(13) des temps creux, il y en a pas dans une maison
'(lit.) Times for slacking off, there aren't any of these in a house.' [026_C-1: 338]

4.1.4 Coordination

In line with the general rule that a finite verb and its arguments constitute one EDU, cases of coordinated finite verbs and finite verb phrases were analyzed as separate EDUs.

(14) en principe (euh) la patronne (euh) coupe | essaie | (euh) certaines couturières coupent | et essaient en dehors (de de le) de l'apprentie (ouv-) ouvrière
'In principle, the boss cuts | tries on | some seamstresses cut | and try (the model) on without the apprentice.' [010_-1: 269–272]

This rule also applies in cases of "gapping", i.e. coordinated clauses where the verb is left out in the second conjunct:

(15) et ces gens-là mariaient leur jeune fille à la fin de juillet | et leur fils le quatre septembre
'And these people married off their young daughter at the end of July | and their son on September 4.' [010_C-1: 343–344]

4.1.5 Eventive noun phrases

Noun phrases can exceptionally form an independent EDU in case they refer to events. An example for an eventive noun phrase is given in (16), where *du fait des bombardements* 'because of the bombing' is semantically equivalent to saying that the place was bombarded, and therefore adds a new proposition to the discourse model.

(16) oui parce que (euh) vous avez une place un peu plus haut (euh) qui a été un peu chamboulée | du fait (de euh) des bombardements
'Yes, because a little higher up there you have a place that was turned upside down | because of the bombing' [010_C-1: 62–63]

4.1.6 Fragments

Not every EDU has to include a finite verb. Fragment answers are analyzed as one complete EDU, as they are fully interpretable in the context of a preceding question.

(17) (JMINT:) qu'est-ce qui vous a amenée à vivre à Orléans alors ?
(1254:) mes parents
'(JMINT:) So what brought you to live in Orleans?
(1254:) My parents.' [026_C-1: 16–17]

Even in the absence of an explicit preceding question, verbless utterances can constitute independent EDUs, especially when a verbal structure is easily recoverable.

(18) ben le samedi soir, pas de sortie du tout
'Well, Saturday evenings, no going out at all.' [013_C-1: 160]

4.1.7 Discourse markers and response particles

Verbal discourse markers, such as *remarquez* 'note', *(vous) voyez* 'you see', *(vous) savez* 'you know', *si vous voulez* 'if you wish', are not analyzed as constituting an independent EDU, but are integrated into the adjacent one. In the same vein, answer particles (*oui* 'yes', *non* 'no'), including phatic response particles (*hm*), are integrated into the adjacent EDU.

In case there is no adjacent EDU by the same speaker, they exceptionally constitute an independent EDU, following the principle that every token must belong to exactly one EDU.

4.2 Attributes

- N: Sequence number of the EDU within the file.

5 Referring expressions (reference)

5.1 Markables

In a first step, phrases with a nominal or pronominal head are considered as potentially referring expressions, i.e. they receive an annotation on the reference layer. This includes noun phrases, names (incl. vocatives), pronouns, and (differently from the original annotation in ANCOR_Centre) possessive determiners. Adverbs are generally not covered in the current stage of the annotation. In a second step, the decision whether or not a potentially referring expression is used referentially is relegated to the REFERENTIALITY attribute.

There is a restricted number of cases where a nominal expression is not marked up as a potentially referring expression. These cases include nominal parts of complex prepositions or conjunctions (*à la suite de*, *à condition que*, *à chaque fois que*, *de façon à ce que*), connectors (*d'une part ... d'autre part*, *d'un côté*), or of fixed adverbial expressions (*en réalité*, *en fait*, *sans doute*, *en tout cas*), where the nominal part can never be taken up anaphorically. In the same vein, directions (*(la) gauche*, *(la) droite*) are not annotated, even when preceded by a determiner. Furthermore, compound nouns, such as *les journaux de mode*, *agent de police*, or *robe de mariée* are analyzed as one referring expression; the bare qualifying noun following *de* is not annotated separately. However, in case they are specified by a determiner, both the complete noun phrase and the qualifying noun phrase are annotated: *les journaux de [la mode]*, *la publicité [du prêt-à-porter]*.

In general, a markable on the `reference` layer spans the entire phrase, i.e. the head of the potentially referring expression and all of its syntactic dependents. One exception to this rule

are relative clauses, where the relative clause is not included in the markable of its nominal or pronominal head. In case the potentially referring expression is part of a prepositional phrase, the preposition is not included in the markable span, except for contractions of preposition + definite determiner (e.g., *du* → *de+le*, *aux* → *à+les*). Note that *de/d'* is included in the markable span when it functions as an indefinite, or partitive, determiner, not as a preposition: *je n'ai pas [d'enfants]* (cf. *vous avez [des enfants],?*), but *je viens d' [Orléans].*

A note on possessives: In the original ANCOR_Centre annotation, possessive expressions did not receive a reference annotation. Here, we assume that possessive determiners do refer. Hence, an expression like *ses enfants* 'his/her children' contains two markables: One spanning the whole expression, referring to the possessee (here: the children), and one spanning only the possessive determiner *ses*, referring to the possessor. The same procedure is adopted for possessive adjectives, e.g. *le sien* 'his/hers (lit.: the his/hers)'.

Null pronouns are annotated when they fill the subject role of a finite verb, as in *… et avait des manches en tulle* in (1). These are the only cases in which verbs are annotated as markables on the `reference` layer. Subjects of non-finite verbs (*PRO* in generative grammar) are annotated only when the non-finite clause constitutes a separate EDU, in order to capture the continuity in reference.

(1) et alors cette robe de mariée était un modèle de chez Patou | et avait des manches (euh) en tulle
'And so, this wedding dress was a model from Patou | and had sleeves made of tulle.'
[010_C-1: 382–383]

5.2 Attributes

5.2.1 Features of the referring expression

An overview of the grammatical and referential features pertaining to referring expressions is given in Table 5.1.

The REFERENTIAL attribute should be set to *idiom* in cases where the potentially referring expression is part of a fixed expression, and thus does not actually refer. Examples for pronouns annotated as *idiom* are: *s'**en** aller*, *(ne pas) **en** pouvoir plus*, ***c'**est **ça***, ***ça** va.* Nominal expressions annotated as *idiom* are nominal parts of light verb constructions, where the verb and the noun together form a non-compositional verbal meaning. Examples are *avoir **l'air***, *faire **attention***, *avoir **besoin***, *prendre **la suite** de quelqu'un.*

Table 5.1: Features of the referring expression

Attribute	Description	Values		*Examples*
PERSON	Grammatical person	*1*	(first person)	*moi, je, me, nous*
		2	(second person)	*toi, tu, te, vous*
		3	(third person)	*lui, elle, l'homme, on*
NUMBER	Grammatical number	*sing*	(singular)	*je, mon école, une fille*
		plur	(plural)	*nous, les enfants, ils*
		undef	(undefined)	*on, y, en, se*
GENDER	Grammatical gender	*masc*	(masculine)	*un lycée, il, chef de section, Jean-Paul*
		fem	(feminine)	*une question, elle, ma fille, Mimi*
		undef	(undefined)	*ça, y, je, tous les deux*
CATEGORY	Morphosyntactic category of the expression	*pron*	(pronoun)	*je, ton, lui, y, en*
		np	(noun phrase)	*les enfants, Marie, une différence*
PRONTYPE	Type of pronoun, only defined for CATEGORY = *pron*	*pers*	(personal pronoun)	*je, tu, il, on, y*
		poss	(possessive determiner or adjective)	*mon, ton, sien son*
		rel	(relative pronoun)	*qui, que, dont, auquel*
		refl	(reflexive pronoun)	*se*
		int	(interrogative pronoun)	*qui, que, où*
		indef	(indefinite pronoun)	*quelqu'un, rien, tout*
		dem	(demonstrative pronoun)	*celui, ça, ce*
		null	(null subjects of finite verbs)	*Ø*

Attribute	Description	Values		*Examples*
NPTYPE	Type of noun phrase, only defined for CATEGORY = *np*	*indef*	(indefinite NP, incl. partitives, quantified NPs and bare nouns)	*un enfant, de l'argent, tous les élèves*
		def	(NP with definite determiner)	*la position, les affaires*
		dem	(NP with demonstrative determiner)	*ce garçon, cette idée*
		poss	(NP with possessive determiner or adjective)	*ses enfants, les miens*
		name	(proper name)	*Jean-Paul, le lycée Pothier, Orléans*
		rel	(NP with relative determiner)	*auquel cas*
		int	(NP with interrogative determiner)	*quel rôle, combien d'étudiants*
SYNROLE	Syntactic role of the expression	*subj*	(subject)	***vous** pouvez parler, **mon père** était fonctionnaire*
		obj	(direct object, incl. predicative objects)	*il y a pas **de répétition**, il est **professeur**, je **l'**ai pas fait*
		obl	(indirect or prepositional object)	*j'habite à **Orléans**, c'est pour **me** faire raconter ma vie, on s'occupe **des enfants***
		mod	(modifier of a verbal head)	*j'ai fait mes études **au lycée Pothier**, **dans ce château**, il y avait deux bouveries*

Attribute	Description	Values		*Examples*
		nmod	(modifier of a nominal head)	*c'est une école d'**ingénieurs électroniciens**, un moine de **St Benoît**, mon père était employé **SNCF***
		det	(possessive determiner)	***sa** mère, **votre** avis*
		conj	(paratactic element, e.g. apposition or non-first conjunct in a coordination)	*ce n'était que des vignes, **des pâturages** et **des arbres fruitiers**,*
		disloc	(dislocated phrases, incl. hanging topics and "double subjects")	*ce sont eux qui vont avoir des bourses, c'est-à-dire **les fils d'agriculteurs*** *et **moi** je suis restée à Orléans,*
		cleft	(clefted phrase)	***la comptabilité**, ça me plaît beaucoup* *ce sont **eux** qui vont avoir des bourses, il y a **les cars** qui passent devant mon bureau*
		other	(all remaining cases)	
REFERENTIAL	Whether the expression is used referentially or not. In case of non-referential uses, the type of non-referential use is specified.	*ref*	(referential)	***j'**ai **deux enfants**, Orléans*
		pred	(predicative use)	*il est **professeur**, vous vous considérez **une artiste** ?*

Attribute	Description	Values		*Examples*
		int	(interrogative expression)	***qui** a dit ça ?, d'**où** est-ce que vous prenez vos étudiants ?*
		expl	(expletive pronoun)	***il y** a, **c'**est moi qui l'ai fait*
		idiom	(idiomatic use, or nominal part of a light verb construction)	*c'est **ça**, appeler **au téléphone***
		neg	(negative quantifiers, or partitives in the scope of negation)	*personne, rien, (pas) d'accident*
		prop	(abstract anaphora, verbal or propositional *ça*)	*je trouve **ça** formidable*
		unclear	(the intended reference cannot be determined)	
INTRO	Whether the expression is used to introduce an entity or not	*intro*	(introduction of new entity)	*J'ai **une fille**. Elle a six ans.*
		not_intro	(non-first mention)	*J'ai une fille. **Elle** a six ans.*

5.2.2 Features of the discourse entity

An overview of the features pertaining to discourse entities is given in Table 5.2.

The ID_ENTITY attribute is filled by an automatically generated ID that is shared by all expressions selected as referring to the same entity (in other words, by all expressions in the same coreference chain). In case there is no other expression marked as co-referent (i.e., the coreference set is a singleton set), the ID_ENTITY attribute is set to *empty* by MMAX2. During the postprocessing of the data, all singleton sets are attributed a unique ID, and existing IDs are made globally unique by prefixed the file name, since MMAX2 IDs are only unique inside the same file.

In general, two expressions are marked as coreferential (as belonging to the same coreference set) when they refer to the same real-world or hypothetical entity. In principle, coreferential expressions should be interchangeable without causing a different interpretation of referential intention of the speaker.

Multiple expressions referring to the discourse participants (usually first and second person pronouns, but also vocatives) are also analyzed as coreferential.

Table 5.2: Features of the discourse entity

Attribute	Description	Values		*Examples*
ID_ENTITY	ID of the discourse entity	{coreference set}		
INFSTAT_INTRO	Information status of the entity upon introduction. Only defined for INTRO = *intro*.	*sit*	(the entity is given in the situational context, e.g. the discourse participants)	*je, moi, vous*
		ident	(uniquely identifiable, either by itself or via information in the precedent context)	*Et puis j'ai travaillé à* ***Paris****, dans* ***le banlieue***
		new	(the entity is brand-new)	*Il y a* ***des tas de cafés****,* ***un dossier*** *était déposé*
ANIMACY	Animacy classification	*human*	(human)	*ma fille, vous, un moine*
		org	(organization)	*la mairie, mon école, Simca*
		animal	(animal)	*les bœufs, un chat*
		place	(potential locations for humans)	*Paris, la banlieue*
		time	(dates, times or duration)	*septembre dernier, dix ans*
		conc	(concrete, i.e. tangible objects)	*mon bureau, une barrière*
		nonconc	(abstract)	*l'isolement, différentes raisons, deux grosseurs différentes*
		veh	(vehicles)	*l'aérotrain, des cars*
SPECIFICITY	Specificity	*spec*	(specific)	*Mademoiselle Pathénatan, mes enfants*
		nonspec	(class, hypothetical or non-specific entity)	*les enfants de salariés, l'industrie automobile*

Indefinite expressions are generally not assumed to be co-referential with a previously established referent. This decision was taken to ensure consistency in the annotation, although some examples emerged where the context strongly suggests a co-referential reading, cf. (2) and (3).

(2) a. (euh hm) moi je vous dirai que personellement (euh je je j'ai t-) j'ai toujours aimé *mon métier*
'Let me tell you that personally, I always loved my profession.' [010_C-1: 178]

b. et puis alors à l'heure actuelle (euh) on a *un métier qui dépérit*
'Also, at this point, we are working in a profession that's dying out.' [010_C-1: 287]

(3) a. je suis partie dans *cette maison* | *qui* était dans le Marais là
'I went to that house, which was located in the Marais.' [010_C-1: 441–442]

b. je suis arrivée dans (un) *une maison avec un fouillis indescriptible*
'I arrived in a house with an indescribable mess.' [010_C-1: 443]

The only exception to this rule concerns reference to classes of entities (i.e., generic reference), which can be achieved either by using a morphosyntactically indefinite or a definite expression. In (4), both *la couturière* and *une couturière* make reference to the same discourse entity, which is the class of seamstresses. They are consequently analyzed as coreferential.

(4) a. (on a l'air de) on a l'air de juger un peu *la couturière* (euh comment vous dirais-je) comme une fille qui n'a rien fait à l'école
'People somehow seem to judge the seamstress, eh, how should I say, as a girl that didn't do anything in school' [010_C-1: 156]

b. alors (euh que quand on est euh) qu'*une couturière*, il faut d'abord qu'elle (sache) sache recevoir ses clientes, (euh) discuter avec sa cliente, (euh) guider sa cliente
'Whereas a seamstress has to know how to receive customers, talk to her customer, guide her customer.' [010_C-1: 159]

In the annotation of information status (INFSTAT_INTRO), we make a basic distinction between entities that are present in the situational context of the utterance, entities that are not present but identifiable, and entities that are new. This attribute is only relevant for referring expressions marked as *intro*, since non-first mentions of an entity are necessarily discourse-given.

Situationally given entities include the discourse participants, as well as references to the here-and-now. Identifiable entities include names as well as definite descriptions, which are assumed to be identifiable by virtue of their lexical content and morphosyntactic marking. All other entities are marked as new.

With regard to ANIMACY, we follow the guidelines proposed by Zaenen et al. (2004). This analysis supersedes the original annotation of named entities, because the linguistic theories to be tested target animacy distinctions regardless of whether the entities are named or referred to using a descriptive NP.[1]

Starting from a three-way distinction between humans, other animates, and inanimates, the tagset introduced by Zaenen et al. (2004) makes a more fine-grained distinction that also integrates differences in concreteness/abstractness as well as borderline cases such as organizations. In addition to the categories shown in Table 5.2, their tagset also included a label *mac* for intelligent machines, which however was not attested in the data and consequently dropped. In the annotation of animacy, the semantics of the head noun alone is not important, but its use in context: In (5), *une place* does not indicate a location in the physical sense, but rather a job position. It is therefore tagged as *nonconc*, not as *place*.

(5) alors comme (euh j'ét-) je n'avais pas une place qui correspondait (à) à ce que j'aurais pu avoir chez Simca | que j'avais essayé de chercher une situation à Paris, enfin, dans différentes branches
'So since I didn't have a position that corresponded to what I could have had at Simca, | I tried to find something in Paris, in different sectors' [004_C-1: 29–30]

[1] For instance, there is a clear difference in animacy between *un enfant* 'a child' and *un bureau* 'an office', which both do not qualify as named entities.

In general, the tag *place* was used only in rather restricted contexts, when the expression referred to a larger area that serves as a location for humans. This is typically the case for geographical locations like countries, regions, cities or neighborhoods. More restricted areas, like offices or houses, were instead analyzed as *conc*.

In this corpus, the distinction between organizations (*org*) and concrete objects *conc* is sometimes difficult to make. We annotate as *org* any institution with a collective purpose (as proposed by Zaenen et al. 2004). This is the case for *Simca*, *SNCF*, but also for *l'industrie automobile* 'the car industrie' and *la mairie* 'the city council'. Furthermore, schools and institutions of higher education were analyzed as *org*. In contrast, entities annotated as *conc* can never depict a group of humans with a collective voice and purpose. Examples for concrete objects are *une maison* 'a house', *le brevet* 'the certificate of secondary education', *des tas de papiers* 'lots of papers'. In addition, body parts such as *la main* 'the hand' are also tagged as *conc*, since they cannot act autonomously.

In the annotation of specificity, we follow the proposal made by Riester & Baumann (2017) (their [± generic])[2]. Specific entities are those that are existent in the real world, and can be identified by the speaker. The label non-specific applies to classes or non-instantiated entities. This is the case for both *une place* and *une situation* in (5) above: In the first case, the negation *je n'avais pas une place qui correspondait à ...* 'I didn't have a position that corresponded to ...' indicates that *une place* is non-instantiated; in the second case, *une situation* can also only be interpreted as non-instantiated because of its role as the object to *chercher* 'to look for'. Questions on the existence of an entity constitute another context for non-specific indefinites, cf. (6), as these entities are potentially non-instantiated. Note that a positive answer (such as *oui, il y a des noms spéciaux* 'Yes, there are special names') would then instantiate the entity. This means that *des noms spéciaux* in the answer would be analyzed as non-coreferential and specific, since in uttering the answer, the speaker confirms the existence.

[2]We do not make a distinction between generic reference, as reference to a whole class of entities, and non-specific entities, which can be individuated hypothetical entities.

(6) il y a des noms spéciaux ?
'Are there special names (for that)?' [010_C-1: 108]

In a similar vein, entities are analyzed as non-specific when they refer to different instantiations over time, or to classes. This is illustrated in (7), where *le bilan* does not refer to a particular balance sheet, but rather expresses the idea that preparing the balance sheets in general is the job of the accountant. An example for reference to classes is *la couturière* in (4) above, as well as *des enfants* and *des parents* in (8).

(7) j'ai un comptable bien sûr | mais enfin qui fait uniquement le bilan
'I have an accountant of course, but (one) who only makes the balance sheet.' [013_C-1: 41–42]

(8) quand il faut retirer des enfants des parents, même des parents mauvais | c'est quelque chose d'épouvantable
'When you have to take children from their parents, even from bad parents, | that's a horrible thing.' [026_C-1: 192–193]

6 Disfluencies and incomplete utterances (`disfluency`)

6.1 Markables

Disfluencies such as fillers, repetitions, false starts and reformulations are marked up as spans on the `disfluency` annotation layer.

The types of disfluencies captured in the annotation include exact repetitions of words, as in (1), but also reformulations of larger spans produced by the same speaker, as in (2) and (3).

(1) et (c'est) c'est beaucoup trop
'And (that's) that's way too much.' [013_C-1: 108]

(2) c'était pour (euh une grosse euh pas de euh) une grosse société automobile (euh) pour la branche commerciale
'That was for a (eh, a big, eh, not a, eh) a big car company, for the commercial branch.' [013_C-1: 87]

(3) il y a un ou deux cafés qui sont ouverts | mais enfin y a pas d'ambiance | (il y a pas, c'est pas encore euh) c'est pas sympa
'There are one or two cafés that are open | but well, there's no atmosphere. | (There's no-, it's not yet-, erm) It's not nice.' [004_C-1: 173–175]

In addition, we mark on the same layer utterances that are left incomplete. This is the case when the speaker was interrupted by another speaker and did not take up the utterance again, as in (4), or, more rarely, the speaker abandoned the utterance just started and decided to continue talking about something else, see (5).

(4) (1254:) parce que (euh) y a bien plus de difficultés après | c'est sûr
(JMINT:) oui oui (euh enfin plus on)
(1254:) y a des essais dans la Seine je crois
'(1254:) Because there are way more problems after, | that's for sure.
(JMINT:) Yeah, yeah (well, the more you-).
(1254:) There are some tests in the Seine I think.' [026_C-1: 373–376]

(5) (faut reconnaître qu'il y a sûrement) d'ailleurs pour le moment (y a) ça progresse maintenant de ce côté-là si vous voulez
'(You have to acknowledge that there's certainly-) Besides, for the moment (there is) it's getting better now in that area if you wish.' [004_C-1: 213]

6.2 Attributes

The disfluency layer currently does not have any attributes.

Bibliography

Carlson, Lynn, Daniel Marcu & Mary Ellen Okurowski. 2003. *Building a discourse-tagged corpus in the framework of rhetorical structure theory* 85–112. Dordrecht: Springer Netherlands. doi: 10.1007/978-94-010-0019-2_5.

Müller, Christoph & Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn & Joybrato Mukherjee (eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods*, 197–214. Frankfurt a.M., Germany: Peter Lang.

Muller, Philippe, Marianne Vergez-Couret, Laurent Prévot, Nicholas Asher, Benamara Farah, Myriam Bras, Anne Le Dracoulec & Laure Vieu. 2012. Manuel d'annotation en relations de discours du projet ANNODIS. Tech. Rep. 21. Institut de recherche en informatique de Toulouse.

Muzerelle, Judith, Anaïs Lefeuvre, Jean-Yves Antoine, Emmanuel Schang, Denis Maurel, Jeanne Villaneau & Iris Eshkol. 2013. ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In *Actes TALN'2013*, 555–563. Les Sable d'Olonne, France.

Muzerelle, Judith, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol & Jeanne Villaneau. 2014. ANCOR_Centre, a large free spoken French coreference corpus: Description of the resource and reliability measures. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC 2014)*, 843–847. Reykjavik, Iceland: European Language Resources Association (ELRA).

Riester, Arndt & Stefan Baumann. 2017. The RefLex scheme – annotation guidelines. *SinSpeC Working Paper of the SFB 732 "Incremental Specification in Context"* 14.

Stede, Manfred, Maite Taboada & Depodam Das. 2017. Annotation Guidelines for Rhetorical Structure. Unpublished manuscript.

Zaenen, Annie, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor & Tom Wasow. 2004. Animacy encoding in English:

Why and how. In *Proceedings of the 2004 ACL workshop on discourse annotation* (DiscAnnotation '04), 118–125. Stroudsburg, PA, USA: Association for Computational Linguistics.