Traitement automatique des langues

États de l'art en TAL

sous la direction de
Cécile Fabre
Emmanuel Morin
Sophie Rosset
Pascale Sébillot

Vol. 63 - n°3 / 2022

ATALA

États de l'art en TAL

ATALA

Revue de
l'Association
pour le Traitement
Automatique
des Langues

TAL
Vol.
63

n°3
2022

États de l'art en TAL

ATALA

Traitement automatique des langues

Revue publiée depuis 1960 par l'Association pour le Traitement Automatique des LAngues (ATALA), avec le concours du CNRS.

ISSN 1965-0906

https ://www.atala.org/revuetal

Traitement automatique des langues

Traitement automatique des langues

Volume 63 – n°3 / 2022

ÉTATS DE L'ART EN TAL

Table des matières

Préface

Ce numéro de la revue TAL inaugure un nouveau type d'appel à soumissions, consacré aux articles proposant un état de l'art dans le champ du traitement des langues.

La possibilité de soumettre un article de ce type était théoriquement ouverte par l'appel Varia (au même titre que les *position papers*), mais peu d'auteurs se saisissaient jusque-là de ce format. L'idée de proposer ce nouveau type de numéro, qui est appelé à être proposé régulièrement par la revue désormais, est née de plusieurs constats. Tout d'abord, la communauté scientifique concernée par la revue est très diverse, confirmant le caractère fortement interdisciplinaire du champ : un questionnaire diffusé cette année sur la liste de diffusion LN a montré que les répondants se répartissaient de la manière suivante : 38 % en informatique, 39,5 % en linguistique informatique, 14 % en linguistique, 8,5 % autres (humanités numériques, éducation, etc.). Les articles publiés par la revue et les intitulés des numéros thématiques sont également, et naturellement, le reflet de cette diversité : citons par exemple, sur les dernières années, les thématiques « diversité linguistique en TAL », « apprentissage profond pour le TAL », « dialogue et systèmes de dialogue », « TAL et humanités numériques ». Il est donc important que ces sous-communautés continuent à partager leurs questions de recherche malgré la diversité de leurs intérêts et de leur expertise. Par ailleurs, le champ du TAL interagit avec des communautés connexes (parole, image, recherche d'information, mathématiques, linguistique, cognition, etc.), qui sont intéressées par les résultats et les méthodes que le TAL produit. Enfin, le champ du traitement automatique des langues a évolué ces dernières années à un rythme très soutenu, nécessitant une mise à jour constante de la part des chercheurs. Tous ces constats ont amené le comité de rédaction de la revue à proposer un numéro dédié aux états de l'art, pour faciliter la circulation et l'appropriation des connaissances en TAL au sein de sa communauté et à destination d'autres communautés scientifiques.

L'appel à soumissions a sollicité des articles rendant compte des travaux les plus importants dans le domaine choisi par les auteurs, et présentant l'évolution de ce domaine pour aboutir aux avancées les plus récentes. La synthèse devait être rigoureuse, claire et accessible pour des lecteurs de la revue non spécialistes de la thématique de l'article. Elle devait proposer une mise en perspective des travaux présentés, permettant de comprendre leur articulation. Les thématiques couvertes étaient celles habituellement ciblées par les numéros Varia de la revue, soit tous les aspects du traitement

automatique des langues écrites, parlées et signées et de la linguistique computationnelle.

Le présent numéro contient les trois articles retenus lors de l'appel non thématique lancé en mars 2022 et clos à la mi-juillet 2022. Huit articles avaient été soumis, soit un taux de sélection de 37,5 %, comparable au taux moyen de la revue sur les dernières années. La moitié des premiers auteurs de ces soumissions initiales étaient issus de laboratoires hors de France. Signalons que l'appel demandait aux auteurs d'informer préalablement de leur intention à soumettre, de façon à permettre aux éditeurs d'anticiper la recherche de relecteurs experts. Dix-neuf résumés ont été envoyés dans cette phase initiale, ce qui, à la fois, indique un intérêt important pour ce type de numéro, et montre la nécessité de mieux calibrer la procédure pour réduire ce décalage entre le nombre d'intentions et le nombre de soumissions effectives. Chaque article, comme il est d'usage dans la revue, a été évalué en double aveugle et expertisé par trois relecteurs, dont deux relecteurs externes spécialistes du champ étudié, et un relecteur interne au comité de rédaction de la revue.

Les trois articles retenus sont les suivants :

– *Discourse annotation - Towards a dialogue system for pair programming*, Cecilia Domingo, Paul Piwek, Svetlana Stoyanchev et Michel Wermelinger (The Open University, Toshiba Europe Limited, UK) : cet article offre une présentation des systèmes de dialogue visant à assister la programmation (*programming-oriented dialog systems*) ;

– *Fillers in Spoken Language Understanding : Computational and Psycholinguistic Perspectives*, Tanvi Dinkar, Chloé Clavel et Ioana Vasilescu (Heriot Watt University, Télécom Paris, Université Paris-Saclay) : cet article propose une synthèse des recherches menées sur les disfluences et leur impact sur la compréhension de la langue parlée, en focalisant l'étude sur le cas des disfluences à un seul token (*filler*) ;

– *Les corpus arborés avant et après le numérique*, Sylvain Kahane et Nicolas Mazziotta (Modyco, Université Paris Nanterre & CNRS, U.R. Traverses, Université de Liège) : il s'agit d'une présentation du rôle et de l'usage des corpus arborés au fil des siècles, depuis l'époque pré-numérique jusqu'à nos jours.

On trouvera à la suite de ces articles des notes de lecture ainsi que des résumés de thèses et habilitations à diriger des recherches récemment soutenues en TAL. Un grand merci à Denis Maurel pour le soin accordé à la préparation de la rubrique *Notes de lecture* au fil des années. Nous encourageons désormais nos lecteurs à contacter Didier Schwab (didier.schwab@univ-grenoble-alpes.fr), qui prend le relai de Denis sur cette rubrique. Pour ce qui est des résumés de thèses ou d'HDR, nous vous invitons à les transmettre à Sylvain Pogodalla (sylvain.pogodalla@inria.fr) afin d'informer notre communauté. Merci à Sylvain pour son travail de veille et de collecte.

Nous remercions les membres du comité de rédaction de la revue qui ont participé aux différentes étapes d'élaboration de ce numéro, et en particulier ceux qui ont pris en charge des relectures (voir la composition du comité sur le site de la revue : `https://www.atala.org/content/comit%C3%A9-de-r%C3%A9daction-0`), ainsi que les

relecteurs spécifiques de ce numéro : Alexandre Allauzen, Jean-Yves Antoine, Frédéric Béchet, Béatrice Daille, Nuria Gala, Nabil Hathout, Nicolas Hernandez, Cyril Labbé, Frédéric Landragin, Gaël Lejeune, Aleksandra Miletic, Ludovic Tanguy, Guillaume Wisniewski.

Nous remercions enfin l'Institut des sciences humaines et sociales (INSHS) du CNRS pour le soutien financier apporté à la revue.

Cécile Fabre
CLLE, Université Toulouse 2
cecile.fabre@univ-tlse2.fr

Emmanuel Morin
LS2N, Université de Nantes
emmanuel.morin@univ-nantes.fr

Sophie Rosset
Université Paris-Saclay, CNRS, LISN
sophie.rosset@lisn.fr

Pascale Sébillot
IRISA, INSA Rennes
pascale.sebillot@irisa.fr

Discourse annotation — Towards a dialogue system for pair programming

Cecilia Domingo* — **Paul Piwek*** — **Svetlana Stoyanchev**** — **Michel Wermelinger***

* *The Open University, United Kingdom*
** *Toshiba Europe Limited, United Kingdom*

ABSTRACT. *Much work has been carried out on dialogue system development in different fields. With recent advances in Programming Language Processing tasks, dialogue systems aimed at programmers are becoming another viable area of application. However, the data necessary for a dialogue system that can assist programmers involves not only code, but the natural language around it. How should this data be annotated? In this review we examine the most common approaches to dialogue annotation, paying special attention to programming settings. We first look at the broader theories that inform these approaches, and after our review of the most widely used annotation schemes we analyze the peculiarities of the programming context and how well suited the existing schemes are for this setting.*

RÉSUMÉ. *Le développement de systèmes de dialogue a fait l'objet d'une grande attention dans différents domaines. Avec les progrès récents des tâches de traitement du langage de programmation, les systèmes de dialogue destinés aux programmeurs deviennent un autre domaine d'application viable. Cependant, afin de développer un système de dialogue pour assister les programmeurs, il est nécessaire de traiter non seulement le code, mais aussi le langage naturel associé. Comment ces données doivent-elles être annotées ? Dans cet article, nous présentons une synthèse des méthodes les plus courantes d'annotation des dialogues, avec un accent particulier sur le domaine de la programmation. On considère d'abord les théories sur lesquelles ces méthodes sont basées, on énumère les principales méthodes et on analyse les particularités du domaine de la programmation et dans quelle mesure les principales méthodes d'annotation sont adaptées à ce domaine.*

KEYWORDS: *dialogue systems, discourse annotation, programming language processing.*

MOTS-CLÉS : *systèmes de dialogue, annotation discursive, traitement du langage de programmation.*

1. Introduction

Dialogue systems have become pervasive in our everyday life, spanning a wide range of domains (Liu *et al.*, 2020; Kuyven *et al.*, 2018; Thoppilan *et al.*, 2022). While programmers put their efforts into developing this wide range of systems, few dialogue systems exist which focus on assisting these programmers. There exist other kinds of tools which assist programmers in some ways. Recently, for instance, Github released Copilot [1], a tool that autocompletes code, and Amazon released the similar CodeWhisperer tool [2]. In educational settings, there are numerous Intelligent Tutoring Systems (ITS) that can give feedback or some guidance to people learning to program (Keuning *et al.*, 2019). However, none of these tools follow a dialogic approach and harness the potential of dialogue systems.

One activity where programmers would highly benefit from a dialogue system assisting them is pair programming. Pair programming is a technique where two programmers work together on one piece of code (Hanks *et al.*, 2011). A sample from a pair-programming session can be found below in Example 1 [3]; we will be using it to illustrate different phenomena. In this session, two programmers demonstrate the technique, writing a program that detects high temperatures from an input list of temperatures. In this particular session, both participants are sitting together in front of the computer. However, only one of the participants takes control of the mouse and keyboard, assuming the role of "driver"; the other participant, the "navigator", collaborates with verbal guidance. The video allows us to see not only the participants, but also the screen where they are coding (albeit with faulty synchronization in this case) and a whiteboard they use for brainstorming, all in different windows merged into the same video. The technique they demonstrate, pair programming, is widely used, and has been proven very beneficial, especially in educational contexts (ibid). However, it can be difficult to implement, due to scheduling problems or partner incompatibility (ibid). A dialogue system could help ameliorate these issues. In fact, Wizard-of-Oz studies have already demonstrated the value of a dialogue system as a pair-programming partner: it could bring many of the benefits of pair programming (e.g., better performance, higher confidence), and it would be valued by users (Kuttal *et al.*, 2021). However, there is not sufficient data to develop such systems (Wood *et al.*, 2018).

Before more data can be made available, it is important to reflect on what this data would look like and how it would need to be processed. Some dialogue systems can be built with unannotated data (Thoppilan *et al.*, 2022). However, such systems require very large amounts of data and high computational power (ibid). A more viable alternative is to follow a supervised approach. Our envisioned system would enter the category of task-oriented systems, as it should collaborate with the human user to

1. https://github.com/features/copilot/.

2. https://aws.amazon.com/codewhisperer/.

3. We recommend watching the video of the session to understand the context of the sample. https://youtu.be/zdE2MS6gcbE.

achieve a goal, which in this case is developing a program. For such dialogue systems, the usual approach is a dialogue-state architecture (Jurafsky and Martin, 2021). These systems have a component that detects relevant entities in an utterance (slot filling), a dialogue state tracker that records the dialogue state at each point by considering the slots and the dialogue acts (which we shall discuss in Section 2.2.1), and a set of dialogue policies that determine the system's actions base don the dialogue state. These components require annotated data for training and testing the components — in a supervised approach, the target slots and dialogue acts need to be known. Then, exactly which annotations do we need for dialogues in order to develop a system to replace a human partner in pair-programming sessions? In this paper we reflect on this issue by analyzing existing theories of dialogue and the annotation schemes derived from them, and compare our findings with the characteristics of pair-programming dialogue.

Example 1 — Sample from pair-programming session

(1)	**Navigator:**	You could, you could put a print list if you want, just to.
(2)	**Driver:**	Okay.
(3)	**Navigator:**	But I would always just run it each time and I generally keep versions as well. I don't think we need to do that here, but I would keep... Could build version A, version B so that, if something goes wrong, you can get back to something that worked.
(4)	**Driver:**	[Typing] Right.
(5)	**Navigator:**	Yeah, again, it's just there, maybe there may be other ways of doing it now that I don't know, but I would just save a, b, c, 1, 2, 3.
(6)	**Driver:**	Okay. [Pointing at the screen] So that's our first version anyway. So saved that.
(7)	**Navigator:**	Just run it.
(8)	**Driver:**	[Overlapping] If we run...
(9)	**Navigator:**	It shouldn't, don't get any errors. Something works. [Looking at screen] Yeah, it works, fine.
(10)	**Driver:**	[Overlapping] There you go. Okay.
(11)	**Navigator:**	Fine. Okay. So where are we? [Looking at reference book] So we've got our input, [looking at whiteboard] so we're back to our pattern, right. So I know it needs a list. Set some sort of variable to the first item in the list.
(12)	**Driver:**	Uhum, okay. So again, uuum, what do we call it, something that's sensible again.
(13)	**Navigator:**	It's the highest temperature, isn't it the term?
(14)	**Driver:**	[Overlapping] [Typing] Aye.
(15)	**Navigator:**	Highest value or something. Put maybe highest temp, it's probably, since you've used temp at the...[Shrugging] Yeah, highest Temp.
(16)	**Driver:**	[Typing] And... we're gonna save that.
(17)	**Navigator:**	[Overlapping] The first item in the list
(18)	**Driver:**	[Typing and mumbling] Sensor...
(19)	**Navigator:**	[Mumbling] Temps...
(20)	**Driver:**	[Typing and mumbling, overlapping] Temps. And... okay.
(21)	**Navigator:**	Yeah.

(22) **Driver:** Wanna do zero?
(23) **Navigator:** Zero.
(24) **Driver:** Zero.
(25) **Navigator:** I think...
(26) **Driver:** Because...
(27) **Navigator:** Because an array always, well, doesn't always, that's the problem. And Python...
(28) **Driver:** [Overlapping, smiling] Python is...
(29) **Navigator:** An array starts...S list, an array. If I'm saying array...amm, a list...
(30) **Driver:** [Overlapping] Yeah
(31) **Navigator:** A list starts at zero.
(32) **Driver:** [Overlapping] Zero.

2. Discourse theories

2.1. *Definition*

Discourse can be defined as "joint activities in which conventional language plays a dominant role" (Clark, 2005, p. 50). This broad definition can be considered even broader if we take Skidmore's continuum of addressivity (Skidmore, 2019), where dialogues can range from the truly dialogic to the monologic. For the purposes of analyzing discourse in relation to dialogue systems, we wish to emphasize in this review the joint-activity aspect of discourse to obtain insights applicable to the more dialogic part of the spectrum (ibid). Thus, we will look at discourse theories as they apply to dialogue and not other types of discourse.

2.2. *Key concepts*

Numerous discourse theories have been developed, with more than a few achieving great influence. Therefore, instead of providing a detailed account of each of them, we will now summarize the key themes they cover.

2.2.1. *Acts and actions*

As we will discuss in more detail in Section 3.2, one of the theories that has had the strongest influence in how dialogue is conceptualized in NLP is Speech Act Theory (Austin, 2018). In this view, utterances are actions performed by the participants in discourse. The theory describes several levels of actions, from the phonetic act of making noises to the perlocutionary act of causing an effect on the hearer.

When the focus of the analysis is on the characteristics of discourse as an action between participants, instead of merely looking at the micro details of phonetics or syntax, we must turn to the level of illocutionary and perlocutionary acts and their effects. Austin (2018) distinguishes numerous types of such effects, such as verdictives (the

act of, as the name suggests, emitting a verdict) or exercitives (the act of issuing a recommendation, like (1) in our example). This classification was then built upon by Searle (1979) and several other authors.

Austin's idea of acts emphasizes the effects of utterances, yet it pays little attention to the interaction between the participants originating and receiving these effects. This cooperative element of the speech acts was explored by Clark and Schaefer (1989).

2.2.2. *Cooperative dimension*

The central view in Clark and Schaefer's (1989) theories is that using language is performing a joint action. Participating in dialogue is not seen as the sum of speakers' individual actions, but as a coordinated activity between them. Perhaps the most influential account of how participants cooperate in discourse are Grice's (1957) maxims: "make your contribution as informative as is required", "try to make your contribution one that is true", "be relevant", "be perspicuous" (Grice, 1991, p. 26). These are the rules that allow speakers to follow the Cooperative principle and achieve the goal of their discourse. The Cooperative principle states that speakers should make the contributions to the joint activity (discourse) that are required to achieve its goal (Grice, 1991). Sperber and Wilson (2010) build on Grice's work to develop their relevance theory. One key aspect of it is that communication relies on participants inferring meaning from the speaker's utterances. Inferences are linked to relevance: the inferred meaning of an utterance should be relevant to the context. Relevance, on its part, is mediated by effort: "an assumption is relevant in a context to the extent that the effort required to process it in this context is small" (*ibid* p. 125). Warren (2006) also builds on Grice's work, observing cooperation as a feature of naturalness in conversation data. A more dialogic equivalent of the Cooperative principle is the principle of least collaborative effort (Clark, 2005): participants will try to minimize the total effort of the joint activity, though this may involve putting additional effort on producing the utterances so that little effort is needed to understand them. For instance, in (5) of our example, the navigator essentially repeats what he says in utterance three, possibly trying to emphasize the goal of the utterances.

Gregoromichelaki *et al.* (2011), addressing some limitations in Grice's theories, discuss another factor that enables cooperation: incrementality. Discourse is produced and processed gradually, which enables participants to adjust it based on the feedback they receive. Yet the most widely discussed concept when studying the social elements of discourse is grounding, or how participants in discourse build a common ground (Clark and Schaefer, 1989; Clark, 2005).

2.2.3. *Common ground*

The common ground in a joint activity can be defined as the shared knowledge available to participants in discourse, be it knowledge about the world or the joint activity itself (Clark, 2005). Joint activities begin with an initial common ground that is built upon as the activity progresses, through accumulation or even deletion (Clark and Schaefer, 1989; Clark, 2005). Conceptualizations of the common ground can be built

as iterative propositions ad infinitum (i.e., it involves the speakers knowing that X is true, knowing that they know, knowing that they know they know, etc.); but actual processing cannot be expected to take place this way (ibid). In conversation, for the common ground to be built upon (grounding), participants need to provide sufficient evidence that they have adequately processed each other's contributions (ibid). This may be done through showing continued attention, making a new contribution that is relevant to their counterpart's, acknowledging understanding, repeating all or part of the contribution or displaying understanding some other way (ibid). While this is still a recursive process, with participants giving evidence of understanding contributions, then of understanding the understanding and so forth, this recursion does not continue ad infinitum, as the required level of evidence becomes weaker for each iteration (ibid). For instance, in our example the driver often acknowledges understanding of the navigator's utterances simply by saying "okay", and then they can move on to a different contribution. The concept of common ground is more concretely conceptualized by Grosz and Sidner (1986), who present the idea of a focus space, a discourse dimension containing the purpose of a discourse segment and the available referents for it. Pickering and Garrod (2004) adopt Clark's definition of the common ground, but argue that most dialogue uses a simpler, implicit common ground. This is built as speakers align their situational models; this alignment facilitates both comprehension and production for the speakers.

Beyond the theory, the concept of common ground has also been empirically tested. For instance, Jordan and Walker (2005) created a model to predict the content of referring expressions, and some of the features they employed reflected the theories on common ground (e.g., previously used attributes, other referents that could act as distractors, attribute saliency, etc.). Although the model's accuracy did not reach beyond 50%, the features related to the focus space showed some predictive power (features related to differences between the referent and distractors present in the focus space). Mitchell *et al*. (2012) also carried out some experiments regarding the common ground through the study of convergence (how participants in discourse adapt to each other). After analyzing tutor-mentee interactions over several weeks, they saw that lexical convergence increased over time: the words preferred by a participant became shared knowledge. An extensive practical study of convergence has also been carried out by Dubuisson *et al*. (2021), who developed a framework to compute measures of alignment and used it to analyze dialogues between humans and between humans and agents, observing differences in the flexibility of the alignment. Their work also opens doors for improving alignment in dialogue systems. Another important contribution that bridges theory and practice is Ginzburg's conversation theory, which builds a grammar for dialogue based on corpus data.

As we have mentioned, the focus space contains the referents and knowledge available and relevant to a particular discourse segment (Grosz and Sidner, 1986). In addition to that, it contains the purpose of the segment (ibid). In Grosz and Sidner's model, that intention that the segment tries to achieve determines how discourse is segmented: discourse units each have their own purpose contribution to the purpose of the overall

discourse (ibid). However, it is not only their theory which is primarily driven by the concept of purpose or intention – intention is another key pillar of discourse theories.

2.2.4. *Intention*

Grosz and Sidner (1986) see intentions as the element that structures discourse: the whole discourse will have a purpose (discourse purpose, or DP), but there will also be sub-goals that define the segments of the discourse (discourse segment purpose, or DSP) and become the key element of the focus space at any point. Even before this model, intention was already seen as a driving force within discourse. Austin (2018) reflected it through the illocutionary force of utterances: people say things to achieve a certain effect, and the type of desired effect is what allowed Austin and then Searle (1975, in Clark, 2005) and other authors to classify utterances. Grice (1991, p.26) also emphasized the importance of intentions through the Cooperative principle: "make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged". Grice (1957) also points out that it is not enough for utterances to have an intention, but that speakers must intend for this intention to be recognized, and that other participants in discourse must recognize the intention. As influential as Grice's theories have been, this and the theories it has guided have not been free of criticism. Gregoromichelaki *et al.* (2011), while accepting the concept of intention as useful for high-level conceptualizations of discourse, reject the claim that it has any significant effect in online processing of discourse due to its intractability. One of their arguments is the fact that people with low skills/experience in detecting other speakers' intentions and emotions (e.g., children and people who have autism but good verbal skills) may be able to participate in discourse (ibid). Instead, what would drive discourse processing is social conventions (ibid). When signals are used in non-conventional ways, there are other resources that can help participants solve the coordination problems of discourse, such as explicit agreement, saliency or precedents (Clark, 2005). For instance, in our example the speakers agree to refer to variables regarding temperature as "temp".

2.2.5. *Structures*

Finally, after having discussed the concepts on which discourse scholars define discourse structure, we can provide a brief account of what these structures look like. We have presented Clark and Schaefer's model (1989) as highly influential; this model, stemming from the idea that discourse is a type of joint action that uses language, views discourse as a tree of linked contributions and acceptances. One participant contributes an utterance, but for it to be added to the common ground and for participants to coordinate the individual actions into the joint action, the contribution needs to be accepted – and the acceptance utterance then becomes a new contribution for the first participant to accept.

Grosz and Sidner (1986) provide another model of discourse structure which accounts for more elements in it. As we mentioned, they establish purpose as the driving force

that defines the discourse segments. The segments can be nested, and dominance and precedence relations can be established between them. These segments with distinct purposes contributing to the overall discourse purpose form the intentional structure of discourse. They also define the attentional structure of discourse: this is a stack of focus spaces, each containing the purpose of the segment and the referents and relations relevant to it. The last structural element of discourse is the linguistic structure, the mere sequence of utterances.

Another type of structure that has been widely studied is rhetorical structure, especially through Rhetorical Structure Theory (RST) (Mann and Thompson, 1988, in Hou *et al.*, 2020). This theory sees the structure of discourse as links between utterances building a coherent unit (Hobbs, 1979, in Moore *et al.*, 2003). In RST, discourse units are schemas consisting of a nucleus and one or more satellites (ibid). The nuclei convey the core content of the discourse, while the satellites support the nuclei or other satellites through rhetorical relations like motivation, generalization, evidence, etc. (ibid). For instance, in (5) of our example, we can observe an antithesis relation between the first and second part. Although RST was initially devised for monologic discourse, it has later been applied to dialogue (Daradoumis, 1993).

2.3. *Links between discourse theories and NLP*

In the previous section we have presented some of the most influential discourse theories, and we have also briefly mentioned some empirical studies that drew from those theories. Still, as Pery-Woodley and Scott (2006) observed, there is some divide between discourse theories and NLP tasks that could benefit from them. Discourse theories help us conceptualize the macro-structure of discourse (e.g., cross-paragraph relations), but NLP tends to be more concerned with the micro-elements (e.g., the purpose of an individual sentence), as macro-structure is more intractable (ibid). On the other hand, discourse theories would also benefit from a link to empirical studies and NLP techniques that could test the theories and find connections between macro- and micro-structures. When it comes to bringing the insights from discourse theories into the NLP domain and, more specifically, to dialogue system development, we identify two key challenges: multimodality and online processing.

2.3.1. *Multimodality in dialogue*

Some of the dominating discourse theories acknowledge that dialogue involves both verbal and non-verbal signals (Clark and Schaefer, 1989; Grice, 1991; Clark, 2005). For instance, in (6) of our example, the driver uses both pointing gestures and demonstrative pronouns to refer to the program. For Clark (2005, p. 13), a signal is "any action by which someone means something for another person", and most combine modalities. Beyond theory, multimodality is a key concept in NLP research, and increasingly so (Admoni and Scassellati, 2014). Thus, we need to examine how truly suitable the theories are to account for the processing of non-verbal signals.

2.3.2. *Online processing*

In dialogue systems, the need for online (i.e., live) processing is undeniable, as "dialogues are created incrementally" (Skantze, 2021, p. 83): when we observe dialogue live, we cannot observe the final product, only its gradual development. Clark (2005, p. 29) already advocated for an "action approach" over a "product approach" to discourse. However, the main theories we have discussed fail to account for incrementality. Grosz and Sidner's model (1986), with the constant updating of the focus space stack, reflects this notion at the utterance level, but does not accurately represent incremental processing of utterance sub-elements (Gregoromichelaki *et al.*, 2011). Skantze (2021) aims to fill this gap with a model of incremental processing focused specifically on dialogue systems.

3. Annotation schemes

Annotation schemes help us bridge the gap between discourse theories and NLP tasks, as they facilitate discourse processing. Two of the main theories that are employed are Rhetorical Structure Theory (RST) and Speech Act Theory, each with a different focus. RST conceptualizes discourse through the relations between units. Although rhetorical relations represent speakers' intentions, a stronger emphasis is placed on the idea of intentions by Speech Act Theory. Another difference between the two theories is that RST emphasizes the relations between discourse units, whereas Speech Act Theory gives more weight to the actual discourse units. Given this divergence, we shall divide our account of annotation schemes into at least two trends.

3.1. *Relational schemes*

3.1.1. *Rhetorical Structure Theory (RST)*

RST conceptualizes discourse as consisting of propositions linked by coherence relations, which give rise to implicit propositions, helping us understand the initial propositions (ibid). In practice, it turns discourse into a set of schemas: a nucleus with satellites linked to it through rhetorical relations. For instance, in (5) of our example, the sentence starting with "but" would be a satellite linked to the rest of the utterance, the nucleus, through an antithesis relation.

Although RST was initially proposed for monologic discourse, it has later also been applied to dialogue. One approach has been to apply it to each turn separately, though this does not account for relations between utterances (Fawcett and Davies, 1992, in Daradoumis, 1993). Daradoumis (1993) addresses this by developing Dialogic Rhetorical Structure Theory (DRST), combining RST with an exchange model. The exchange model classifies RST schemas through additional dialogic relations: consent, elicitation, ascertainment. Within the schemas, it incorporates Clark and Schaefer's theory (Clark and Schaefer, 1989) by marking contribution and support relations. Another difference between DRST and RST is its dynamic nature: dialogue itself is

dynamic, so the representation is constantly changing, with schema nuclei and satellites changing their status (Daradoumis, 1993).

RST has successfully been harnessed for different NLP tasks, most notably summarization and text generation (Taboada and Mann, 2006; Hou *et al.*, 2020). Some research has also been done in dialogue: e.g., Fischer *et al.* (1994) design a dialogue system for database queries, where RST allows the system to represent the links between dialogue acts. NLP research using this theory has largely been enabled through the Rhetorical Structure Theory Discourse Treebank (RST-DT) (Carlson *et al.*, 2003). It is a corpus of 385 *Wall Street Journal* articles annotated by humans with 78 RST relations (ibid). Another key resource for NLP tasks involving discourse relations is the Penn Discourse Treebank (PDTB) (Prasad *et al.*, 2019), the largest resource of its type. Like the RST, it consists of annotated *Wall Street Journal* articles, but for the PDTB the number is 2,159 (Hou *et al.*, 2020).

Segmented Discourse Representation Theory (SDRT) SDRT was developed to address the shortcomings of RST and Discourse Representation Theory (DRT) (Lascarides and Asher, 2008). DRT (Kamp, 1981, in Asher & Lascarides, 2008) uses first-order logic to build semantic representations of discourse; combining these representations with RST reduces anaphora resolution options to the more "pragmatically preferred" (Lascarides and Asher, 2008, p. 1).

3.1.2. *Mapping relational annotation schemes*

With at least three very influential schemes for annotating discourse relations (RST, SDRT, PDTB), attempts have been made to map them for higher resource reusability. One approach is finding an intersection (Demberg *et al.*, 2019), or another is using an additional schema (Sanders *et al.*, 2021; Roze *et al.*, 2019). Bunt and Prasad (2016) proposed an ISO standard for coherence relation annotation; they also mapped it to the PDTB and RST.

3.2. *Speech Act Theory*

As discussed in Section 2.2, Speech Act Theory (Austin, 2018) sees utterances as the performance of speech acts. Acts range from the mere phonetic act of making sounds to the illocutionary act of uttering something with a purpose and the perlocutionary act of uttering something that has an effect (ibid). For instance, (15) of our example (Section 1) has the effect of the driver typing the variable name that the navigator suggested. The function of an utterance is called the illocutionary force, which can be implicit or can be made explicit through illocutionary force indicating devices (ifids) (Allan, 1997). These can be words like "please", the use of the imperative mood (as in utterance 15), or illocutionary verbs (verbs that perform an action by being uttered, such as "to name"). Austin focused on this last resource for the classification of illocutionary forces. Searle (1979) expanded upon Austin's classification and pointed out that speech acts can be implicit: the illocutionary force of an utterance

may be hidden behind the apparent performance of a different illocutionary act, such as when commands are softened by phrasing them as questions. Speech Act Theory has been critiqued on the basis that it focuses on illocutionary forces (what the speaker intends with an utterance), but these are not always evident: not to discourse annotators, and often even to hearers and speakers (Allan, 1997; Gregoromichelaki *et al.*, 2011). Another issue is that Speech Act Theory looks at speech acts after they are completed, whereas speakers process them incrementally in real time (ibid). Yet another critique is that Speech Act Theory looks at utterances as isolated acts (Allwood, 1977; Allan, 1997), without considering how they contribute to joint activities (Clark and Schaefer, 1989). Moreover, Austin and Searle's focus on illocutionary verbs places too much emphasis on the lexical dimension of discourse, neglecting others (Allwood, 1977). Despite all the criticism, Speech Act Theory has become highly influential for dialogue annotation, with numerous schemes based on it.

3.2.1. *Schemes derived from Speech Act Theory*

A large number of annotation schemes have been created which label speech acts (also known as "dialogue acts" in this context). For instance, Klein *et al.* (1999) already discussed sixteen influential schemes. Among the speech-act labelling schemes, the most widely used seem to be DAMSL and HCRC MapTask (either directly or through adapted versions).

DAMSL This scheme was developed by Allen and Core (1997) with the name Dialogue Act Markup in Several Layers. It has also been augmented for use with the Switchboard corpus (Jurafsky and Shriberg, 1997). As the name suggests, this scheme captures utterance intentions in several layers. The first layer, the forward communicative functions, concerns the speech acts (illocutionary acts). The other two layers are the backward communicative functions (how the utterance relates to the previous one), and the utterance features (form and content of the utterance). Unlike Speech Act Theory as developed by Austin (2018) and Searle (1979), DAMSL assigns several labels to an utterance (in addition to the scheme having three layers, each utterance can have several labels for each layer); this way it can capture the speech act with more nuance. This schema uses very generic labels with the goal of being applicable to any domain (e.g., statements are either assertions or reassertions; if not, they are simply "other"). However, this can result in some broad labels dominating in a corpus: for instance, Stolcke *et al.* (2000) annotated the general-domain Switchboard corpus and assigned the Statement label to more than a third of utterances. Weisser (2015) aimed to address this shortcoming of DAMSL by expanding it into a more nuanced scheme: DART. Whereas DAMSL has fewer than 15 functions per layer, DART has a total of 120 functions in its second version (Core and Allen, 1997; Weisser, 2015). Despite the large number of labels, it is intended to be simple to use thanks to its clear XML structure, which is aimed for easy automation and human interpretation (Weisser, 2015). DART has already been applied to a wide variety of dialogue types, such as political debates, call-center interactions or courtroom dialogue (ibid).

HCRC MapTask The HCRC MapTask project aimed to obtain dialogues that could be manipulated for some linguistic features (e.g., phonological characteristics of landmarks mentioned) (Anderson *et al.*, 1991). It collected data using a cooperative task design named the "map task", where two participants each have a map, but only one of them has information on the route that the other needs to follow on their map (Brown *et al.*, 1984). This approach to data gathering, which allowed for controlling numerous variables, as well as the corpus itself and the way it was annotated, have been highly influential. DAMSL, more closely linked to Speech Act Theory, focuses on the annotation of individual speech acts, paying only some limited attention to the relations between them. The annotation scheme of the HCRC MapTask, on the other hand, has three levels of annotations that go from the macro-structure of dialogue to the individual acts (Kowtko *et al.*, 1993; Carletta *et al.*, 1997). The highest level consists of transactions, which are sub-dialogues that serve to achieve the purpose of the overall dialogue. The middle level consists of dialogue games: they start with an initiation move and end when the goal of the game is achieved or abandoned. Finally, the lower level consists of the actual moves, a limited set of twelve speech acts. While the tagset of moves and games is widely used (Kowtko *et al.*, 1993; Chen and Di Eugenio, 2013; Ribeiro *et al.*, 2022), annotation at the level of transactions is very difficult, resulting in low inter-annotator agreement (Carletta *et al.*, 1997).

Adapted schemes It is difficult to find the right degree of granularity in a scheme so that the nuances of dialogue are accurately captured, but without making annotation too complex. For instance, one problem with studies that use versions of DAMSL is that a large number of utterances are labelled as "statement", so no distinction can be made between them (Margolis *et al.*, 2010; Robe *et al.*, 2020). The need for additional layers and tags is even more evident in tasks that involve more than textual data. Arguably, though, all NLP tasks could benefit from processing input in more than one modality: Stolcke *et al.* (2000), for instance, achieve greater speech act classification accuracy when combining word features with prosodic features. Some tasks involve yet more modalities, such as processing hand gestures. One example is the ELDERLY-AT-HOME multimodal corpus, where researchers had to expand the HCRC MapTask tagset to include labels that could describe the haptic actions involved in the task of providing care to elderly people — e.g., grabbing an object that the other person asks for (Chen and Di Eugenio, 2013).

3.3. *ISO standard*

As we have described, numerous annotation schemes exist, some focused on coherence relations, some more focused on speech acts, all with different degrees of granularity, and with varied suitability for multimodal annotation. In order to promote resource reusability, an ISO standard has been proposed for dialogue annotation, derived from the DIT++ annotation scheme (Bunt, 2009; Bunt, 2019; ISO, 2019). It appears to be mostly influenced by Speech Act Theory, as the focus is on annotating utterance function (ibid). It is intended to be suitable for multimodal annotation. Though

the annotation guidelines do not define any specific tags for non-textual modalities (e.g., eye gaze, facial gestures, hand movements, etc.), contributions in these modalities could be tagged with any of the standard's communicative function labels which described the function of the gesture. Furthermore, Petukhova and Bunt (2012) give some detailed examples of how such contributions could be coded following the ISO standard. In our example, for instance, the annotations for utterance 6 would feature a label called "handMove" with a shape parameter set to "pointing". The standard aims to be highly flexible: it consists of hierarchical labels across two general-domain functions and nine domain-specific functions (Bunt, 2009; Bunt, 2019; ISO, 2019). Depending on the needs of annotators, labels can be assigned on as many functions as is seen fit, and the hierarchy can be employed to the depth that is deemed appropriate (ibid). For instance, Zarisheva and Scheffler (2015) used the ISO standard on Twitter conversations; they attempted to simplify annotation by assigning labels on only one function per utterance, but found that this lowered inter-annotator agreement. The hierarchy also provides flexibility not only for how deeply or widely it can be used, but also for its expansion: the task function is given empty in the standard, leaving it to researchers to complete it with whichever functions they require (Bunt, 2009; Bunt, 2019; ISO, 2019). The ISO standard also defines a representation standard, the Dialogue Act Markup Language, which offers many options for enriching the annotations (ibid). For instance, it supports adding rhetorical relations through another ISO standard; thus, this standard can combine the two main trends in discourse annotation, with both speech act and rhetorical relation labels (ibid). In addition to the rhetorical relations, the standard also defines its own dependence relations between utterances: feedback and function relations (ibid). The standard has already demonstrated its usability for NLP tasks. For instance, Ribeiro *et al*. (2022) successfully apply convolutional neural networks to the classification of speech acts in dialogue text annotated with the standard.

3.4. *Segmentation*

We have discussed different ways in which dialogue segments can be labelled. However, the question remains of how to divide dialogues into relevant segments. In relational schemes, the common practice seems to be segmenting utterances by clauses (elementary discourse units) when using RST, and by sentences when using the PDTB scheme (Poláková *et al.*, 2017; Sanders *et al.*, 2021). When following a scheme based on Speech Act Theory, one option is annotating full turns (i.e., everything a speaker says before being interrupted by another, or before ending the dialogue) (Das and Pon-Barry, 2018). However, a common approach is annotating by utterances, which can be equivalent to a turn, but also to a set of turns or a turn fragment (Bunt, 2009; Bunt and Prasad, 2016; Bunt, 2019). What does then define the boundaries of an utterance? Core and Allen (1997), in line with Grosz and Sidner's (1986) theory of intentional structure, find these boundaries in the changes of function: when the speech act starts fulfilling a new function, that marks the start of a new segment. Nevertheless, Weisser (2015) acknowledges that this concept of "utterance" is often very vaguely defined;

"utterance" is then defined by Weisser as the smallest independent unit with semantic and pragmatic content. Nonetheless, most works on dialogue annotation fail to provide an account of how segmentation was carried out, leaving the definition of utterance and segment vague and/or implicit.

3.5. *Processing dialogue without annotating discourse*

Discourse theories have been highly influential for the creation of annotation schemes to build the corpora used for developing dialogue systems. Nonetheless, such systems can also be built without complex discourse data. In fact, Pery-Woodley and Scott (2006) point out that the macro-structures of discourse are built through processing micro-structures, but macro-structures are not linguistically explicit, which is one of the main issues for discourse processing in NLP. An option then is to focus on the smallest micro-elements. This is viable, for instance, in medical dialogue systems: Liu *et al*. (2020) build a dialogue system that identifies patients' gastrointestinal problems based on the recognition of entities (specific words describing symptoms and medicines). Advances in deep learning and the higher computational capacity that has enabled these advances also offer the alternative of developing dialogue systems without annotating discourse. Recently, for example, Thoppilan *et al*. (2022) trained transformer models with billions of unannotated dialogues and other texts to develop a generic dialogue system. The models were then fine-tuned with data obtained by having crowd-workers interact with the system (ibid).

4. Pair-programming discourse

As we have detailed in the previous section, there are numerous schemes available for dialogue annotation. Some of them are focused on specific tasks, whereas others aim to be applicable to a wide range of dialogue types; some analyze discourse through the relations between segments, whereas others look at the speaker's intention when uttering the segments. And there is also the ISO standard, which combines the different approaches and claims to be suitable for a wide range of dialogue annotation tasks (Bunt, 2009; Bunt, 2019; ISO, 2019). Are this and other schemes then well-suited for the annotation of dialogues produced in programming tasks, or how would they need to be customized? To answer this question, we first need to look at what distinguishes dialogue in this context from other types of dialogue.

We can observe several different types of interactions among programmers – there are even instances of programmers talking to a cardboard image of another programmer to develop their ideas (Bryant *et al.*, 2008). We would like to focus on pair-programming interactions, as this is a widely employed and studied practice which has proven to be very beneficial (Hanks *et al.*, 2011). It has also been shown that a dialogue system facilitating this type of interaction would be highly valuable (Robe *et al.*, 2020; Kuttal *et al.*, 2021; Robe, 2021). However, no such tool exists yet – annotated pair-programming dialogue would bring its development closer to reality.

Such data is lacking; in fact, even unannotated pair-programming data is scarce. Most works on pair programming are primarily concerned with the results of the practice, so researchers perform their analysis on the code produced or on participant surveys rather than dialogue samples (Werner and Denning, 2009; Hanks *et al.*, 2011; Adeliyi *et al.*, 2021). There are some exceptions; below we summarize the main characteristics of pair-programming dialogue that we have been able to extract from the literature. As we will detail in Section 5.1, we later aim to gather our own data to enrich the pool of knowledge on this type of dialogue. In Section 1, we already provided a brief definition of pair programming, as well as a sample transcription from a session. It is part of the Agile approach to software development; it is "a technique in which two individuals share a single computer as they work together to develop software" (Hanks *et al.*, 2011). The programmers normally take either of two roles: the driver and the navigator – though there is some debate about whether these roles are so clearly distinct or even always present (Hanks *et al.*, 2011, p. 135). The driver is the person in charge of the mouse and keyboard, who writes the code, while the navigator offers some guidance. These roles are often switched. Below we detail what the roles imply for the dialogue, as well as other characteristics of pair-programming dialogue. We then end this section with a brief illustration of the phenomena that are observable in the sample we used as an example.

4.1. *Roles*

Some of the literature on pair programming, especially when offering guidelines for the practice, defines very distinct navigator and driver roles, and recommends switching frequently. However, there often is no observable difference in how navigator and driver talk (Bryant *et al.*, 2008) with regard to levels of abstraction; the main difference is mainly control of the keyboard, which may affect how decisions are made, unless both participants have exactly equal access to the keyboard – such as when there are two keyboards. Still, differences may be observed in the amount of talk, as the driver will be focused on typing. Below we summarize how the roles affect some features of the dialogue.

Switching roles: Switching frequency might be imposed, especially in educational settings. However, students may find this hinders the natural workflow (Tsompanoudi *et al.*, 2013). On the other hand, in some sessions the roles are fixed, particularly if participants do not have two keyboards to switch easily. When switches happen, they're often taking advantage of pauses; sometimes, though, the navigator requests to switch. Sometimes it is easier for the navigator to switch and type than to explain what they mean (Chong and Hurlbutt, 2007; Zarb and Hughes, 2015).

Driver "muttering": The drivers, while typing, may verbalize what they are typing; as most of their attention may be devoted to the typing, their verbalization may be in the form of "muttering" (Zarb and Hughes, 2015). Whatever the form of the verbalization, it is encouraged by experts, as it allows the navigator to understand what the driver is thinking and see whether they need assistance (ibid).

Navigator giving instructions: Navigators are encouraged to offer suggestions to contribute to the program (ibid). In less equitable interactions, these suggestions may be given in a more domineering tone (Lewis and Shah, 2015). However, such commands without justification hinder collaboration (Wegerif and Mercer, 1996).

Driver deciding unilaterally: Another example of inequitable interaction and, thus, unsuccessful collaboration, is when the driver makes decisions on their own. When only one keyboard and mouse is available for pair programming, whoever has control of it has the final say about what makes it into the code (Chong and Hurlbutt, 2007). For minor decisions, it may be easier for the navigator to accept what the driver chose instead of arguing (ibid).

4.2. *Multimodality*

Like any other form of spoken dialogue, pair programming involves more than verbal communication (Clark, 2005). Additionally, as pair programming involves working on some code, the code will be commonly referenced and thus become a key element of the discourse.

Spoken discourse: The fact that we are dealing with spoken dialogue has some implications for its structure. The goal of an utterance may not be clear to the speaker from the beginning (Gregoromichelaki *et al.*, 2011), and this may be reflected in the structure: sentences may be ungrammatical, there may be repetitions, speakers may stutter, etc. This is also true for the overall structure of the discourse: regardless of each individual speaker's intentions, the intentional structure of the discourse arises from the joint dialogue (Grosz and Sidner, 1986). Spoken discourse also introduces prosodic features as an important modality to analyze. In unstructured discourse like a natural pair-programming interaction, utterances may be as simple and ambiguous as "mmmm"; a sound like this can indicate both approval and disapproval depending on the tone (Zarb and Hughes, 2015). Prosody also helps speakers manage turns (Skantze, 2021).

Gestures: Facial expressions and other body movements play an important role in communication. For instance, eye gaze can signal attention and serve as a turn-management device (Heylen *et al.*, 2002). A special kind of gesture is pointing (Chen and Di Eugenio, 2013); this gesture allows speakers to call a referent into the discourse. In the case of pair programming, this can be reference materials (a book, a whiteboard, a digital guide, etc.) or an element of the code. Additionally, pair programming has the peculiarity that pointing can be performed with either the body or an input device.

Actions replacing utterances: Non-verbal actions can contribute to the joint activity just like linguistic actions (Clark, 2005). The driver coding as a reaction to a suggestion from the navigator is a form of uptake (ibid). Also, sometimes the navigator may find it easier to request to drive and type in code than having to describe it (Chong and Hurlbutt, 2007).

4.3. *Skill levels*

When the participants have different skills levels, that has an impact on the collaboration (Chong and Hurlbutt, 2007; Plonka *et al.*, 2015). Due to the interaction of diverse factors, pairing programmers with different levels may result in both equitable and inequitable collaboration (Lewis and Shah, 2015). The goal of the session must be borne in mind: it may be knowledge transfer between expert and novice, or it may be simply to finish a project (Chong and Hurlbutt, 2007). Where time pressure is more important than the need to learn, the expert may take a more dominant role, giving direct instructions or driving without giving explanations (ibid). The novice, on their part, may take a passive role to avoid feeling like they are slowing down project completion or making a fool of themselves in front of the expert (ibid). When priority is given to knowledge transfer, an expert navigator may use mentoring strategies ranging from least to most explicit: hinting at problems, pointing out specific problems, and giving clear explanations (Plonka *et al.*, 2015). When the expert is driving, they can still mentor the novice by verbalizing their thoughts (ibid).

4.4. *Domain*

Pair programming is a specialized task and, as such, requires specialized terminology. This can come both from the programming domain and from the domain of the real-world problem that the code aims to solve. The combination of these domains also means that the content of the dialogue will switch among different "levels of abstraction": from the concrete level of the programming language's syntax, to the abstract real-world problem, going through the intermediate level of discussing code sections (Bryant *et al.*, 2008). It has been hypothesized that each of the two pair-programming roles deal with different levels. However, studies have contradicted this hypothesis, showing that both participants speak and initiate dialogue segments at all levels of abstraction (Chong and Hurlbutt, 2007; Bryant *et al.*, 2008) (as measured through specific coding schemes that tag these labels). The studies have also shown that most talk occurs at the intermediate level of abstraction, the discussion of code sections (ibid). Dialogue may also occur outside any of the levels, for instance, when the programmers deviate from the task and talk socially (Bryant *et al.*, 2008). However, these off-topic utterances may also be valuable for the interaction: e.g., they may help programmers disconnect from a difficult problem and reapproach it with a fresh perspective (Zarb and Hughes, 2015). Other disruptions to the workflow, of course, may be undesirable interruptions. One such kind may be an interruption from a third party outside the pair, or the pair programmers getting distracted (Chong and Siino, 2006).

4.5. *Collaboration*

Pair programming is a collaborative task, as participants work towards solving the same task jointly. For collaboration to be successful, participants need to discuss the

task with each other, make joint decisions, build on each other's ideas, correct each other's mistakes, and justify their contributions (Wegerif and Mercer, 1996; Bigman *et al.*, 2021). These features can be observed in successful implementations of the pair-programming technique. A good understanding between pair-programming partners is known as "jelling" (Adeliyi *et al.*, 2021). Jelling can result in some peculiar discourse phenomena. For example, participants may repeat what the other said to show uptake. Other phenomena resulting from good rapport may make the dialogue difficult to interpret to third parties. For instance, when there is great common ground, it may not be necessary for participants to finish their sentences: they may be finished by their partner or be left unfinished.

4.6. *Example*

In the example that we have presented (Section 1), we can observe many of the features that we have discussed. The participants only had one keyboard, which may have deterred them from switching roles. They have a similar skill level, which may have facilitated their successful collaboration. They establish such a good connection that they are even able to finish each other's thoughts, evidence of great shared common ground (Clark, 2005): e.g., when they discuss array characteristics in Python. Despite the equitable collaboration, we are able to see some of the characteristics that distinguish the role of driver and navigator: e.g., the navigator offers suggestions, and the driver types them in, in one instance verbalizing what they are typing through "muttering". The driver's actions also remind us of the importance of multimodality: non-linguistic actions such as saving a file when the navigator suggests doing so is a form of uptake. Multimodality is also important for turn-taking: we can clearly see the participants facing each other to indicate the end of a turn, especially after questions. They also often point at the screen or reference materials when discussing them. We also see several features reflective of spoken discourse. For instance, some sentences are left unfinished (e.g., "I think..."). We also see the navigator starting to discuss how arrays always start a certain way, only to change his mind mid-sentence (turn 27). Another interesting feature is the large number of sentences starting with "so" (e.g., turns 7, 11, or 12), structuring the discourse in an improvised manner. In this fragment we cannot see the whole range of abstraction levels that can be discussed in a pair-programming session — here we mainly see the intermediate level (discussing broad aspects of the program), which may be the most frequent one in pair programming (Chong and Hurlbutt, 2007; Bryant *et al.*, 2008). What we observe abundantly, though, is the use of terminology: the participants are often referring to temperatures (problem domain terminology), and using many programming terms (e.g., "print list", "variable", "array", etc.).

5. Conclusions

In the previous sections, we have looked at the most influential discourse theories and dialogue annotation schemes, as well as the main characteristics of dialogue during pair-programming sessions; thus, we may now attempt to answer our initial question: how can pair-programming dialogue be annotated for its analysis and the development of NLP-based systems that can assist programmers in these sessions?

Discourse theories, particularly Clark and Schaeffer's work (Clark and Schaefer, 1989), have allowed us to see that dialogues are joint activities. As such, they consist of contributions that need to be accepted by partners (uptake). Especially in a context like pair programming, where we seek effective collaboration, we need to annotate whether this uptake takes place. The ISO standard (Bunt, 2009; Bunt, 2019; ISO, 2019) captures this concept through dependence relations between utterances (feedback and functional relations), as well as feedback and dialogue management functions.

Grosz and Sidner's (1986) influential description of an attentional space and Clark's (2005) theories of common ground also show us the value of representing the speakers' shared basis. The ISO standard (Bunt, 2009; Bunt, 2019; ISO, 2019) and all the schemes based on Speech Act Theory allow us to annotate the purpose of segments; SDRT (Lascarides and Asher, 2008), on the other hand, allows us to build a semantic representation of the segments and find the most pragmatically plausible referents to resolve anaphora. With segment purposes and referents, we have the main elements of the focus space described by Grosz and Sidner (1986). As we mentioned in Section 3.1, SDRT combines DRT and RST. While the ISO standard does not include the logic representations of DRT, it does allow for annotation of the coherence relations of RST.

Another important aspect highlighted by influential scholars like Clark (2005), and evident also from observations of pair-programming sessions, is that multimodality is very important. The ISO standard (Bunt, 2009; Bunt, 2019; ISO, 2019) is designed to allow for the annotation of non-verbal contributions to discourse. The standard functions may not accurately describe all non-verbal contributions in pair programming, but the scheme allows for customization, especially through the Task dimension (ibid). For instance, we observed that sometimes a contribution may simply be the driver coding. Additionally, we observed pointing as a frequent form of non-verbal contribution. Chen and Di Eugenio (2013) adapt the HCRC MapTask scheme to add labels to describe pointing and other similar hand gestures; a similar Gesture dimension might be added to the ISO standard. The standard also allows for the optional annotation of three types of qualifiers: certainty, conditionality, and sentiment. These labels could allow annotators to capture the information conveyed through prosody and other non-verbal modalities.

We have mentioned the Task dimension of the ISO standard as an option for labelling task-related functions of non-verbal contributions, such as the driver coding. This dimension could code a lot of valuable information for the analysis of pair programming. For instance, Wood *et al*. (2018) distinguish utterances not simply by general

function, like Statement or Question, but also by programming-related topic (e.g., API or implementation). Other studies code stages of programming-related problem solving, such as "reviewing code", "muttering while typing", and "suggesting" (Zarb and Hughes, 2015). These studies also see the value of off-task contributions; less desirable interruptions of the workflow may also occur frequently, and as such might need to be coded to be differentiated (Plonka *et al.*, 2012). Another crucial concept related to pair programming is the participant's role. While it may not always be distinguishable from looking at the utterances (Chong and Hurlbutt, 2007; Bryant *et al.*, 2008), it would be useful to code who is using the keyboard and/or who has access to it. Roles are already annotated in some corpora in other settings, such as the HCRC MapTask corpus, which annotates the roles of Giver (the person giving directions) and Follower (the person following directions) (Anderson *et al.*, 1991).

Lastly, as the goal in pair programming is often for programmers to improve their skills through collaboration, it might be valuable to add labels stemming from pedagogical theories. For example, a very influential code for dialogue in collaborative learning tasks was developed by Wegerif and Mercer (1996). These authors distinguish one type of talk, disputational talk, which is not conducive to learning, as it merely fosters conflict. Tsan *et al*. (2021) offer a more fine-grained classification of conflict that includes some positive conflict. Task-related conflict may even constitute what Wegerif and Mercer (1996) consider the most effective kind of talk for learning: exploratory talk. Werner and Denning (2009) incorporate Wegerif and Mercer's theories with other codes inspired by Vygotskyan theories. Useful codes are also found in Plonka *et al*. (2015), who analyze mentoring strategies.

5.1. *Future work*

With this work, we aimed to draw some conclusions about how pair-programming dialogues could be annotated for the development of dialogue systems. To do this, we have extracted the most relevant insights from the literature. We started by looking at the theoretical work, to then examine more practical work on dialogue annotation and on the analysis of pair programming. This last section described our conclusions, linking the characteristics of dialogue in general, and pair-programming dialogue in particular, with existing annotation schemes. However, we have put special emphasis on the ISO standard. Standards can make linguistic resources easier to compare and reuse. The standard for dialogue annotation claims to be usable for a wide range of dialogue types and annotation uses, and it can be customized. Our analysis of the literature leads us to suggest that it might be suitable for the annotation of pair-programming dialogue, provided some customization is made and some of the optional annotations are used, as discussed above in Section 5. A Gesture dimension would need to be added, and the Task dimension would need to be developed with labels related to coding actions, pair-programming roles and features of collaboration. The optional sentiment qualifiers would be useful, and dependence relations should also be annotated.

Pair-programming dialogue and similar dialogues have been annotated with simpler schemes (Robe *et al.*, 2020; Kuttal *et al.*, 2021; Robe, 2021), but this may result in most of the dialogue being fit into one category. The imbalanced categories then make it difficult to train models that can be usable for NLP tasks, such as intent detection in dialogue systems (ibid): how can a dialogue system choose the appropriate policies if all it knows for most utterances is that they are statements, without any further distinction? The higher granularity that we recommend here, on the other hand, also has disadvantages. Firstly, complex annotation schemes may result in low annotation reliability. Additionally, the annotation process becomes slower and thus more expensive. Therefore, our next work will be testing our hypothesis that a customized version of the ISO standard is suitable for pair-programming data before any large corpus can be annotated. We have applied the insights from this paper to annotate the video from which we extracted our Example 1 (Section 1). This has allowed us to make some initial decisions about our annotation guidelines, which is the recommended procedure before annotating a larger corpus (Fuoli, 2018). Over the next months, we are planning to annotate additional videos that we have obtained; this shall help us better analyze pair-programming dialogue, as well as continue to refine our annotation scheme by introducing additional annotators. Afterwards, we plan to collect our own data to be able to gather richer multimodal input. Once our data is collected, we aim to start training and testing a subsystem of an NLU component for slot filling (see Section 1). It is our hope that the outcome of our research will then pave the way for the development of further system components and eventually the development of a dialogue system that can effectively function as a pair-programming partner.

* ACKNOWLEDGEMENT: This work has been carried out with financial support from EPSRC Training Grant DTP 2020-2021 Open University and Toshiba Europe Limited.

6. References

Adeliyi A., Wermelinger M., Kear K., Rosewell J., "Investigating Remote Pair Programming In Part-Time Distance Education", *United Kingdom and Ireland Computing Education Research conference*, ACM, Glasgow United Kingdom, p. 1-7, 2021.

Admoni H., Scassellati B., "Data-Driven Model of Nonverbal Behavior for Socially Assistive Human-Robot Interactions", *Proceedings of the 16th International Conference on Multimodal Interaction*, ACM, Istanbul Turkey, p. 196-199, 2014.

Allan K., "Speech Act Theory: An overview", *Concise encyclopedia of philosophy of language*, Pergamon, Exeter, UK, p. 454-467, 1997.

Allwood J., "A Critical look at Speech Act Theory", *Logic, Pragmatics and Grammar*, University of Göteborg, Lund, Sweden, p. 53-99, 1977.

Anderson A., Thompson H. S., Bader M., Bard E., Boyle E. H., Doherty-Sneddon G., Garrod S. C., Isard S. D., Kowtko J. C., McAllister J., Miller J., Sotillo C. F., Weinert R., "The HCRC Map Task Corpus: A Natural Spoken Dialogue Corpus", *Language and Speech*, vol. 34, nº 4, p. 351-366, 1991.

Austin J. L., *How to do things with words: The William James Lectures deliveres at Harvard University in 1955*, Martino Fine Books, Eastford, CT, 2018.

Bigman M., Roy E., Garcia J., Suzara M., Wang K., Piech C., "PearProgram: A More Fruitful Approach to Pair Programming", *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, ACM, Virtual Event USA, p. 900-906, 2021.

Brown G., Anderson A., Shillcock R., Yule G., *Teaching talk*, 1st edn, Cambridge University Press, 1984.

Bryant S., Romero P., du Boulay B., "Pair programming and the mysterious role of the navigator", *International Journal of Human-Computer Studies*, vol. 66, nº 7, p. 519-529, 2008.

Bunt H., "The DIT++ taxonomy for functional dialogue markup", *8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, Budapest, Hungary, 2009.

Bunt H., *Guidelines for using ISO standard 24617-2*, Technical report, Tilburg University, 2019.

Bunt H., Prasad R., "ISO DR-Core (ISO 24617-8): Core Concepts for the Annotation of Discourse Relations", *ACL 2016*, 2016.

Carletta J., Isard A., Isard S., Kowtko J., Doherty-Sneddon G., Anderson A., "The Reliability of a Dialogue Structure Coding Scheme", *Computational Linguistics*, vol. 23, nº 1, p. 13-31, 1997.

Carlson L., Marcu D., Okurowski M. E., "Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory", *in* N. Ide, J. Véronis, J. van Kuppevelt, R. W. Smith (eds), *Current and New Directions in Discourse and Dialogue*, vol. 22, Springer Netherlands, Dordrecht, p. 85-112, 2003.

Chen L., Di Eugenio B., "Multimodality and Dialogue Act Classification in the RoboHelper Project", *SIGDIAL*, Metz, France, p. 183-192, 2013.

Chong J., Hurlbutt T., "The Social Dynamics of Pair Programming", *29th International Conference on Software Engineering (ICSE'07)*, IEEE, Minneapolis, MN, USA, p. 354-363, 2007.

Chong J., Siino R., "Interruptions on software teams: a comparison of paired and solo programmers", *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work — CSCW '06*, ACM Press, Banff, Alberta, Canada, p. 29, 2006.

Clark H. H., *Using language*, 6. print edn, Cambridge University Press, Cambridge, 2005.

Clark H. H., Schaefer E. F., "Contributing to Discourse", *Cognitive Science*, vol. 13, nº 2, p. 259-294, 1989.

Core M., Allen J., "Coding Dialogs with the DAMSL Annotation Scheme", *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, Cambridge, MA, p. 28-35, 1997.

Daradoumis T., "Towards a Representation of the Rhetorical Structure of Interrupted Exchanges", *Fourth European Workshop on Trends in Natural Language Generation, An Artificial Intelligence Perspective*, p. 106-124, 1993.

Das R., Pon-Barry H., "Turn-Taking Strategies for Human-Robot Peer-Learning Dialogue", *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, Association for Computational Linguistics, Melbourne, Australia, p. 119-129, 2018.

Demberg V., Scholman M. C., Asr F. T., "How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations", *Dialogue & Discourse*, vol. 10, nº 1, p. 87-135, 2019.

Dubuisson Duplessis G., Langlet C., Clavel C., Landragin F., "Towards alignment strategies in human-agent interactions based on measures of lexical repetitions", *Language Resources and Evaluation*, vol. 55, n° 2, p. 353-388, 2021.

Fischer M., Maier E., Stein A., "Generating Cooperative System Responses in Information Retrieval Dialogues", *Proceedings of the 7th International Workshop on Natural Language Generation*, Kennebunkport, Maine, 1994.

Fuoli M., "A stepwise method for annotating appraisal", *Functions of Language*, vol. 25, n° 2, p. 229-258, 2018.

Gregoromichelaki E., Kempson R., Purver M., Mills G. J., Cann R., Meyer-Viol W., Healey P. G. T., "Incrementality and intention-recognition in utterance processing", *Dialogue & Discourse*, vol. 2, n° 1, p. 199-233, 2011.

Grice P., "Meaning", *The Philosophical Review*, vol. 66, n° 3, p. 377-388, 1957.

Grice P., *Studies in the Way of Words*, 1991.

Grosz B., Sidner C., "Attention, intentions, and the structure of discourse", *Computational Linguistics*, vol. 12, n° 3, p. 175-204, 1986.

Hanks B., Fitzgerald S., McCauley R., Murphy L., Zander C., "Pair programming in education: a literature review", *Computer Science Education*, vol. 21, n° 2, p. 135-173, 2011.

Heylen D., van Es I., Nijholt A., van Dijk B., "Experimenting with the Gaze of a Conversational Agent", *International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*, Copenhagen, Denmark, p. 93-100, 2002.

Hou S., Zhang S., Fei C., "Rhetorical structure theory: A comprehensive review of theory, parsing methods and applications", *Expert Systems with Applications*, vol. 157, p. 113421, 2020.

ISO, ISO/DIS 24617-2, Second Edition, Technical report, 2019.

Jordan P. W., Walker M. A., "Learning Content Selection Rules for Generating Object Descriptions in Dialogue", *Journal of Artificial Intelligence Research*, vol. 24, p. 157-194, 2005.

Jurafsky D., Martin J., "Chatbots & Dialogue Systems", *Speech and Language Processing*, Stanford University, p. 1-39, 2021.

Jurafsky D., Shriberg E., "Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual", `https://web.stanford.edu/~jurafsky/ws97/manual.august1.html`, 1997. Accessed on March 10th, 2023.

Keuning H., Jeuring J., Heeren B., "A Systematic Literature Review of Automated Feedback Generation for Programming Exercises", *ACM Transactions on Computing Education*, vol. 19, n° 1, p. 1-43, 2019.

Klein M., "An Overview of the State of the Art of Coding Schemes for Dialogue Act Annotation", *in* G. Goos, J. Hartmanis, J. van Leeuwen, V. Matousek, P. Mautner, J. Ocelíková, P. Sojka (eds), *Text, Speech and Dialogue*, vol. 1692, Springer Berlin Heidelberg, Berlin, Heidelberg, p. 274-279, 1999.

Kowtko J. C., Isard S. D., Doherty-Sneddon G., "Conversational Games Within Dialogue", *HCRC Technical Report*, vol. 31, p. 1-12, 1993.

Kuttal S. K., Ong B., Kwasny K., Robe P., "Trade-offs for Substituting a Human with an Agent in a Pair Programming Context: The Good, the Bad, and the Ugly", *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ACM, Yokohama Japan, p. 1-20, 2021.

Kuyven N. L., André Antunes C., João de Barros Vanzin V., Luis Tavares da Silva J., Loureiro Krassmann A., Margarida Rockenbach Tarouco L., "Chatbots na educação: uma Revisão Sistemática da Literatura", *RENOTE*, 2018.

Lascarides A., Asher N., "Segmented Discourse Representation Theory: Dynamic Semantics With Discourse Structure", *in* H. Bunt, R. Muskens (eds), *Computing Meaning*, Springer Netherlands, Dordrecht, p. 87-124, 2008.

Lewis C. M., Shah N., "How Equity and Inequity Can Emerge in Pair Programming", *Proceedings of the 11th annual International Conference on International Computing Education Research*, ACM, Omaha Nebraska USA, p. 41-50, 2015.

Liu W., Tang J., Qin J., Xu L., Li Z., Liang X., "MedDG: A Large-scale Medical Consultation Dataset for Building Medical Dialogue System", *arXiv:2010.07497 [cs]*, 2020.

Margolis A., Livescu K., Ostendorf M., "Domain Adaptation with Unlabeled Data for Dialog Act Tagging", *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, Association for Computational Linguistics, Uppsala, Sweden, p. 45-52, 2010.

Mitchell C., Boyer K. E., Lester J., "From strangers to partners: examining convergence within a longitudinal study of task-oriented dialogue", *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Seoul, South Korea, p. 94-98, 2012.

Petukhova V., Bunt H., "The coding and annotation of multimodal dialogue acts", *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey, p. 1293-1300, 2012.

Pickering M. J., Garrod S., "Toward a mechanistic psychology of dialogue", 2004.

Plonka L., Sharp H., van der Linden J., "Disengagement in pair programming: Does it matter?", *2012 34th International Conference on Software Engineering (ICSE)*, IEEE, Zurich, p. 496-506, 2012.

Plonka L., Sharp H., van der Linden J., Dittrich Y., "Knowledge transfer in pair programming: An in-depth analysis", *International Journal of Human-Computer Studies*, vol. 73, p. 66-78, 2015.

Poláková L., Mírovský J., Synková P., "Signalling Implicit Relations: A PDTB - RST Comparison", *Dialogue & Discourse*, vol. 8, n° 2, p. 225-248, 2017.

Prasad R., Webber B., Lee A., Joshi A., "Penn Discourse Treebank Version 3.0", `https://catalog.ldc.upenn.edu/LDC2019T05`, 2019. Accessed on March 10th, 2023.

Péry-Woodley M.-P., Scott D. R., "Computational Approaches to Discourse and Document Processing", *Trait. Autom. des Langues*, vol. 47, p. 7-19, 2006.

Ribeiro E., Ribeiro R., Martins de Matos D., "Automatic Recognition of the General-Purpose Communicative Functions Defined by the ISO 24617-2 Standard for Dialog Act Annotation", *Journal of Artificial Intelligence Research*, 2022.

Robe P., *"Designing a Pair Programming Conversational Agent"*, Master's thesis, University of Tulsa, Tulsa, Oklahoma, 2021.

Robe P., Kaur Kuttal S., Zhang Y., Bellamy R., "Can Machine Learning Facilitate Remote Pair Programming? Challenges, Insights & Implications", *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, IEEE, Dunedin, New Zealand, p. 1-11, 2020.

Roze C., Braud C., Muller P., "Which aspects of discourse relations are hard to learn? Primitive decomposition for discourse relation classification", *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Association for Computational Linguistics, Stockholm, Sweden, p. 432-441, 2019.

Sanders T. J., Demberg V., Hoek J., Scholman M. C., Asr F. T., Zufferey S., Evers-Vermeul J., "Unifying dimensions in coherence relations: How various annotation frameworks are related", *Corpus Linguistics and Linguistic Theory*, vol. 17, n° 1, p. 1-71, 2021.

Searle J. R., "A taxonomy of illocutionary acts", *Expression and Meaning: Studies in the Theory of Speech Acts*, Cambridge University Press, p. 1-29, 1979.

Skantze G., "Turn-taking in Conversational Systems and Human-Robot Interaction: A Review", *Computer Speech & Language*, vol. 67, p. 101178, 2021.

Skidmore D., "Dialogism and education", *The Routledge International Handbook of Research on Dialogic Education*, p. 27-37, 2019.

Sperber D., Wilson D., *Relevance: communication and cognition*, 2nd edn, Blackwell, 2010.

Stolcke A., Ries K., Coccaro N., Shriberg E., Bates R., Jurafsky D., Taylor P., Martin R., Ess-Dykema C. V., Meteer M., "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech", *Computational Linguistics*, vol. 26, n° 3, p. 339-373, 2000.

Taboada M., Mann W. C., "Applications of Rhetorical Structure Theory", *Discourse Studies*, vol. 8, n° 4, p. 567-588, 2006.

Thoppilan R., De Freitas D., Hall J., et al, "LaMDA: Language Models for Dialog Applications", *arXiv:2201.08239 [cs]*, 2022.

Tsan J., Vandenberg J., Zakaria Z., Boulden D. C., Lynch C., Wiebe E., Boyer K. E., "Collaborative Dialogue and Types of Conflict: An Analysis of Pair Programming Interactions between Upper Elementary Students", *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, ACM, Virtual Event USA, p. 1184-1190, 2021.

Tsompanoudi D., Satratzemi M., Xinogalos S., "Exploring the effects of collaboration scripts embedded in a distributed pair programming system", *Proceedings of the 18th ACM conference on Innovation and technology in computer science education — ITiCSE '13*, ACM Press, Canterbury, England, UK, p. 225, 2013.

Warren M., *Features of naturalness in conversation*, J. Benjamins, 2006.

Wegerif R., Mercer N., "Computers and Reasoning Through Talk in the Classroom", *Language and Education*, vol. 10, n° 1, p. 47-64, 1996.

Weisser M., "Speech act annotation", *in* K. Aijmer, C. Rühlemann (eds), *Corpus Pragmatics*, Cambridge University Press, Cambridge, p. 84-114, 2015.

Werner L., Denning J., "Pair Programming in Middle School: What Does It Look Like?", *Journal of Research on Technology in Education*, vol. 42, n° 1, p. 29-49, 2009.

Wood A., Rodeghero P., Armaly A., McMillan C., "Detecting Speech Act Types in Developer Question/Answer Conversations During Bug Repair", `http://arxiv.org/abs/1806.05130`, 2018. Accessed on March 10th, 2023.

Zarb M., Hughes J., "Breaking the communication barrier: guidelines to aid communication within pair programming", *Computer Science Education*, vol. 25, n° 2, p. 120-151, 2015.

Zarisheva E., Scheffler T., "Dialog Act Annotation for Twitter Conversations", *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Association for Computational Linguistics, Prague, Czech Republic, p. 114-123, 2015.

Fillers in Spoken Language Understanding: Computational and Psycholinguistic Perspectives

Tanvi Dinkar* — **Chloé Clavel**** — **Ioana Vasilescu*****

* *Heriot Watt University, T.Dinkar [at] hw.ac.uk*
** *Télécom Paris, chloe.clavel [at] telecom-paris.fr*
*** *University Paris-Saclay, ioana [at] limsi.fr*

ABSTRACT. *Disfluencies are ubiquitous to spoken discourse. Fillers ("uh", "um", ...) occur the most frequently compared to other kinds of disfluencies. Yet, to the best of our knowledge, there isn't a resource that brings together the research perspectives influencing Spoken Language Understanding (SLU) on these speech events. The aim of this article is to synthesise a breadth of perspectives in a holistic way; i.e. from underlying (psycho)linguistic theory on fillers, to their annotation and consideration in Automatic Speech Recognition (ASR) and SLU systems, to lastly, their study from a generation and Text-to-Speech (TTS) standpoint. The article aims to present the perspectives in an approachable way to the SLU and Conversational AI community, and discuss what we believe are the trends and challenges in each area.*

KEYWORDS: *Disfluencies, Fillers, Spoken Language Understanding.*

RÉSUMÉ. *Les disfluences sont omniprésentes dans le discours, et les fillers ("euh", ...) sont le type de disfluence le plus fréquent. Pourtant, il n'existe aucune ressource qui rassemble les perspectives de recherche sur ces événements discursifs dans le cadre de la compréhension de la langue parlée (CLP). L'objectif de cet article est de synthétiser un large éventail de perspectives de manière holistique, comprenant la théorie (psycho-)linguistique des fillers, leur annotation et leur prise en compte dans les systèmes de reconnaissance automatique de la parole et de CLP, ainsi que leur étude dans le cadre de la génération. L'article a pour but de présenter ces perspectives de manière accessible à la communauté de la CLP et des systèmes conversationnels, et de discuter de ce que nous voyons comme les tendances et défis de chaque domaine.*

MOTS-CLÉS : *Disfluences, Fillers, Compréhension de la Langue Parlée*

1. Introduction

Speech production is a complex process; i.e. "respiratory, phonatory, and articulatory gestures are timed in such a way as to produce an acoustic signal that adequately conveys the intended message both quickly and smoothly" (Lickley, 2015). *Disfluencies* can be thought of as elements that break this fluent flow of speech during speech product ion. Formally, they are interruptions in the regular flow of speech, such as pausing silently, repeating words, or interrupting oneself to correct something said previously (Fraundorf *et al.*, 2018). They occur between intentional signals (a gesture used to point at an object, ...) and *un*intentional signals (slips of the tongue, ...) of communication (Corley and Stewart, 2008). With the increasing popularity of voice assistant technologies, an open challenge that remains is the ability to design systems that can comprehend the different signals of speech communication. However, when taking stock of a machine's capability to understand language, there is an emphasis on the learning of *forms* (such as in Language Modelling (LM), where the task is string prediction), but not on *meaning*; or the relationship between linguistic form and communicative intent (Bender and Koller, 2020). This issue is now widely acknowledged, for instance, the UnImplicit workshop [1] state as motivation "... an important question that remains open is whether such methods are actually capable of modelling how linguistic meaning is shaped and influenced by *context*, or if they simply learn superficial patterns that reflect only *explicitly* stated aspects of meaning ...". From a research standpoint, this is especially challenging when considering disfluencies, which can have *implicit* and *contextual* meanings – such as when a speaker says "uh", "er" and so on. Hence, this work is motivated by the present observations:

People rarely speak in the same manner with which they write. As Bailey and Ferreira (2003) state, "The processes involved in speaking and in writing differ substantially from each other, and so the products of the two systems are not the same". The following is an example of a transcription taken from a corpus of conversational speech:

> **A:** For a while there I, I, I, uh, subscribed to *New York Times*, a-, actually a couple of newspapers because, uh, you know, my fiance, well, she was unemployed for a while ...
> **B:** Uh-huh.
> **A:** ... so she, you know, really needed to look at the, the, want-, help wanted ads.

The above transcript is not *fluent*, and not easily *readable*. Already, some modifications have been done to the raw transcript keeping in mind the goal of readability. For instance, punctuation markers have been added to introduce sentence structure and capture the prosodic cues used by the speaker. In the transcripts, we can see that people tend to repeat themselves ("I, I, I"), interrupt each other ("Uh-huh"), rapidly shift the focus of the topic in a conversation (from newspaper subscriptions to unemployment), and are in general, *disfluent*. One of the departures from written text

1. https:unimplicit.github.io.

is the presence of disfluencies. Disfluencies are frequent in speech, as fluent speech is rarely the norm; with estimates that natural human-human conversations comprise of $\approx 5 - 10\%$ of disfluencies (Shriberg, 1994). *Thus disfluencies are ubiquitous to spontaneous speech.*

Despite this, there are varying attitudes in the treatment of disfluencies depending on the field, that influence research in Spoken Language Understanding (SLU). Generally, the field of psycholinguistics (dealing with communicative and cognitive aspects of language) focuses on the role disfluencies play in the production and comprehension process of speech, with many works to show their importance in speech communication. For instance, they inform us about the linguistic structure of an utterance: such as in the (difficulties of) selection of appropriate vocabulary while circumventing interruption. Other perspectives, such as a *computational* one, focus on recognising disfluencies in order to remove them – for the improvement of automatic speech recognition (ASR) systems.

Furthermore, "disfluencies" is an umbrella term used to describe a wide variety of communicative phenomena. It is difficult to find an overarching definition of disfluency due to several confusions in terminology. Lickley (2015) categorises all taxonomies of disfluencies as based on *form* or *function*. Terminology to describe disfluencies may be based on form when the objective is to describe the "patterns of words and syntactic units that disfluencies display". On the other hand, function descriptions may be used when disfluencies are described with respect to the planning processes involved in speech production, deliberating over the reasons for a departure from fluency (though overlap exists between the two categories). For instance, "uh" may be described as an interjection (word or phrase independent of grammatical connections to other words or phrases), or a "filled pause"/ "filler" (sound filling a pause in flow of speech) when considering form, or a "hesitation" when considering function. When "uh" is used as a (meta-) "discourse marker" (a unit of talk that brackets speech (Schiffrin, 1987)), it considers both form and function – i.e. functions to mark discourse structure, and occurs at discourse boundaries (form). Other examples of functional disfluencies can include "self"/ "other-repair" ("repair" specifically used to indicate that something has gone wrong in the planning process, and needs to be corrected). For more details of annotation used for disfluency detection tasks (which is typically form-based), please refer to Sec. 3.1 Annotation for Disfluency Detection.

This work is concerned with fillers ("uh", "um" ...), compared to other disfluencies. These speech events may not contribute to the final message when considering purely a lexical level (Vasilescu *et al.*, 2010), and thus do not have *explicit* meaning (Meteer *et al.*, 1995). Yet, they occur with *high frequency* in speech datasets, compared to other structures of disfluencies. Shriberg (2001) shows that the number of fillers per word across corpora exceed other kinds of disfluencies. Additionally, they occur in an *intersection between speech and text*. Fillers are a common property of spontaneous speech and are shown to have *distinct acoustic characteristics/ paralinguistic properties* (Shriberg, 1999; Vasilescu *et al.*, 2010), yet they can be transcribed in text, without the requirement of more detailed annotation schemes. Thus, with the

growing demand of voice assistant technologies, there is an increase in research that deals with spontaneous speech corpora – where naturally, *these tokens will frequently occur*.

Yet so far, there isn't a body of work that brings together the research relevant to SLU on these specific disfluencies. To the best of our knowledge, there is no *introductory work* that brings together several perspectives surrounding fillers, i.e. from psycholinguistic to computational. While a linguistic versus computational perspective is discussed in other works (Lickley, 1994), there is a narrow focus within this perspective on disfluency detection for the purposes of removal. This ignores the broader applications of SLU that deal with *social communication*. Additionally, while other linguistic works give rigorous details about the annotation of disfluencies (Nicholson, 2007; Grosman, 2018), including the history behind these schemes (Lickley, 1994), they are not accessible for the SLU community. This work is more targeted towards researchers in SLU and dialogue (or indeed, linguists interested in the work on fillers and disfluencies in SLU). Given the recent benchmark release of a filler detection task and dataset (Zhu *et al.*, 2022), there is a growing interest in the community on these unique disfluencies.

Thus, **the aim of this paper** is to bring together a *breadth* of perspectives on these spontaneous speech phenomena in a holistic way, considering fillers at different stages in a pipeline; i.e. from ASR to generation. Furthermore, we draw parallels in the fields of psycholinguistics and SLU; discussing research in both fields that considers fillers informative, noise .... Throughout the work, we discuss the challenges in each field; from issues in safety and robustness, to a lack of contextual analysis and feasibility in in-the-wild scenarios. To do so, we contrast computational approaches to disfluencies distinguished from psycholinguistic approaches, loosely adopted from Lickley (1994). Broadly, psycholinguistic theories deal with the communicative and cognitive aspects of language. For instance, Levelt (1983) studies how speakers monitor and correct their speech, and in turn, how listeners are able to integrate new material correcting the previous material. Thus in Sec. 2 we focus on psycholinguistic research that studies the role disfluencies in the *planning/production* of speech, and the *comprehension* of speech. As discussed, we focus on the research surrounding fillers; highlighting works that study them as informative signals of communication. However, we also present research on other disfluencies when relevant – this is to discuss general concepts and findings from the field. Then, in Sec. 3, we discuss the computational perspectives on disfluencies; i.e. approaches more concerned with the *recognition/processing* of disfluencies. Here, we discuss how the treatment of disfluencies can vary depending on the *type of task* in SLU. Sec. 4 then gives the conclusion of the article.

This paper can be read non-linearly by referring to the legend given in Table 1. It is difficult to create a one-to-one mapping in the fields, given that the methodologies and approaches to each field are different. However, the intention behind some works are similar, and we believe that they can be grouped together as complementary perspectives.

	Psycholinguistic	**Computational**
Production	2.1.1 Cognitive load 2.1.2 Communicative Function	3.2.1 Broader SLU
Comprehension	2.2.1 Informative cues 2.2.3 Results on incremental processing 2.2.4 Time-buying measures	3.2.2 Generation
Noisy channel	2.2.2 Filtered out noise	S3 (intro) Computational Perspectives S3.1 SLU for SDS
Note on annotation S1 Introduction	X	3.1.3 Annotation for Disfluency Detection

Table 1. *Legend to show the sections of each perspective that can be read as complementary to the other.*

2. Psycholinguistic Perspectives: from Production to Comprehension

In this section, we discuss works that study the planning/production of speech by the speaker, and the comprehension of speech by the listener [2]. Please note, we consistently utilise the term "fillers" in this work, including when citing previous research [3] that may use other terms to describe the same phenomena [4].

2.1. *Production*

2.1.1. *Cognitive load*

A common theme in the planning process is the idea of *cognitive load*, i.e. the amount of cognitive *effort* required in the planning process. The production of disfluencies resulting from the planning process has been considered at different linguistic levels. For example, a prosodic/acoustic analysis found that speakers tend to main-

2. We utilise both terms "planning" *and* "production", because it is not clear how intentional or unintentional disfluencies are as signals uttered by the speaker (Nicholson, 2007).

3. Please note, there are several other linguistic perspectives not discussed in this paper. Since disfluencies are ubiquitous to spontaneous speech, the literature on disfluencies is vast. For example, a socio-linguistic focus; such as studying the effect that gender, regional background, etc. have on the the production of disfluencies in Shriberg (2001). Some of these perspectives, including works from psycholinguistics, *do consider disfluencies as noise*. We refer the reader to the works of Nicholson (2007) and Lickley (1994), which give a comprehensive overview of these perspectives.

4. Clark and Fox Tree (2002) introduced the term "fillers"; as the term "filled pauses" seemed to indicate that there is a pause in speech *filled* by some sort of (meaningless) sound.

tain a fixed speaking rate during most utterances, but often adopt a faster or slower rate, depending on the cognitive load (O'Shaughnessy, 1995). From this, it was found that disfluent speech may be a result of cognitive load; i.e speakers slow down their speech when having to make unanticipated choices (for example, using more fillers), and accelerate their speech when repeating some words. This shows that the *types of disfluencies produced* may give further insight to this planning process. The link between the rate of speech and the different disfluencies produced was also found in Shriberg (2001). Two groups of speakers were identified – *repeaters*, who produce more repetitions (when what was said is exactly repeated) than deletions (when previous material that was uttered is abandoned), and *deleters*, who produce more deletions than repetitions. The *repeater-deleter* difference was not only due to stylistic variation in speakers; deleters have a faster speaking rate than repeaters in terms of words per unit time. The interpretation suggested here (contrary to O'Shaughnessy (1995)) is that speakers with a slower speaking rate (repeaters) "take more time to plan", leading to an increase in repetitions, while faster speakers (deleters) "get ahead of themselves", and recant what was said to begin again. The different conclusions could be due to the *difference in dataset size, context and domain* [5]. For instance, speakers have been shown to be more disfluent in dialogues compared to monologues (Oviatt, 1995), in human-human conversations than human-machine conversations (Oviatt, 1995), and disfluencies are affected by dialogue role and domain (Colman and Healey, 2011).

The analysis of disfluencies at other levels also reveal *characteristics of the planning process*. In an acoustic-syntactic analysis, it was found that high-frequency monosyllabic function words (such as "the" or "I") are more likely to be prolonged or have a fuller form when there are neighbouring fillers ("uh" and "um"), indicating that the speaker was encountering problems in planning the utterance (Bell *et al.*, 2003). At an utterance level, Shriberg (2001) found that the longer the utterance, the more disfluencies they contain, also suggesting an increase in cognitive load of the speaker. Disfluencies were also found to occur at the start of an utterance, due to higher cognitive load in planning an utterance (Maclay and Osgood, 1959). At a discourse level, Swerts (1998) found that fillers can be used by the speaker to indicate pausing to (re)formulate thoughts, particularly at discourse boundaries.

2.1.2. *Communicative function*

Beattie and Butterworth (1979) suggested that instead of cognitive load, there is an *element of speaker choice* in the planning process. They first establish that generally, speakers are disfluent both when producing low frequency words and improbable words in the context, focusing on fillers "ah", "er", and "um". However, even when frequency of a word was maintained by the speaker, they were still disfluent when producing words with low contextual probability. Like this, there are two main positions behind the speaker's production of disfluencies. One is that disfluencies are acciden-

5. Indeed, Shriberg (2001)'s findings were based on conversational style dialogues (human-human) in addition to task-oriented dialogues, while O'Shaughnessy (1995) focused on task-oriented dialogues (human-machine).

tally caused in speech due to *cognitive burden* of the speaker (such as Bard *et al.* (2001)). Other works study disfluencies as an important *communicative function* used in dialogue. This view is based on the *strategic modelling view*, where the speaker strategically updates the listener, by using disfluencies as cues (Nicholson, 2007). The distinction between the two views is that the former is an unconscious by-product of speech produced by cognitive (over)load, while the latter is an intentional and strategic production by the speaker. Often studies will look at both of these positions, by analysing the individual disfluencies of a speaker as well as the collective disfluencies produced by interlocutors. The results for the production of disfluencies are often mixed, such as in Nicholson (2007) and Yoshida and Lickley (2010), with evidence suggesting that they occur in both cases; i.e. speakers may both strategically and unconsciously produce disfluencies depending on the context.

2.2. *Comprehension*

Research also focuses on the *comprehension* of disfluent speech, i.e. taking into account the listener's understanding of the speaker's disfluencies (Corley and Stewart, 2008), and not on why the disfluency itself was produced (Nicholson, 2007). As Corley and Stewart (2008) state, "it is hard to determine the reason that a speaker is disfluent, especially if the investigation is carried out after the fact from a corpus of recorded speech". Works that study the effect of disfluencies on listener comprehension state that listeners *must* have developed a comprehension system to process disfluencies, given how frequent they are in spoken language.

2.2.1. *Informative cues*

Research shows that listeners can use disfluencies as signals to *understand and resolve incoming information* in the flow of speech, regardless of whether the speaker intentionally used disfluencies in that way. Consider the following example taken from Brennan and Williams (1995):

A: Can I borrow that book?
B: . . . {F um} . . . all right.

Here, speaker **B** used a filler {F...}, which causes **A** to note that **B** might have had a different intention compared to if **B** answered "all right" immediately. While this example is on a pragmatic level (i.e. the listener notes the discrepancy between what was said in essence, and how it was said), works also suggest that listener's can use disfluencies as *communicative cues* in other ways.

For instance, it was found that fillers helped in the faster recognition of a target word for listeners, indicating that they cause listeners to pay more attention to the upcoming flow of speech (Fox Tree, 1995). The use of fillers also *biases* listeners towards new referring expressions rather than ones already introduced into the discourse (Arnold *et al.*, 2004). Arnold *et al.* (2007) additionally showed that listeners have expectations on the upcoming material to contain *difficult to describe*/unconventional

referring expressions when preceded by the filler "uh". Listeners expect the speaker to refer to something new following the filler "um", compared to noise of the same duration (such as a cough or sniffle) (Barr and Seyfeddinipur, 2010). This result was found to be *speaker specific* for the listener. This means that the listener was able to consider what was an already introduced referring expression versus a new referring expression for the current speaker; not just what was old or new for themselves (the listener). Barr and Seyfeddinipur (2010) suggest that this is evidence for the *perspective taking account of language comprehension*. This account suggests that listeners are able to interpret fillers as delay signals, and then infer on plausible reasons for this delay in speech by considering the speaker's perspective. This implies that fillers can have a *metacognitive (i.e. assessment of knowledge state) effect*, with the listener using fillers as cues to interpret the speaker's *metacognitve state*. Fillers and prosodic cues were also found to impact listener's attributions of a speaker's metacognitive state, specifically the estimation of a speaker's level of certainty on a topic (Brennan and Williams, 1995).

2.2.2. *Filtered out noise*

As a counterpoint to the works discussed, research also suggests that disfluencies may be perceived as noise and thus filtered out by the human listener. In experiments, Fox Tree (1995) noted that removing repetitions from the utterance digitally did not affect the rating of perceived naturalness of speech. However, this could be explained from later work by Shriberg (1999), who found that repetitions have similar pitch contours, but just stretched out over time. Lickley *et al.* (1991) found that listeners could not pinpoint the exact interruption point of a disfluency, and tended to point it out later (up to one word) in the flow of speech. Lickley and Bard (1998) studied a listener's ability to identify (several types of) disfluencies to find that they are not reliably predictable unless a noticeable pause or abandoned word is apparent. Note, the works discussed here on disfluencies as cues to integrate information are mainly targeted towards fillers, compared to these works that study more complex disfluencies [6].

2.2.3. *Results on incremental processing*

Bailey and Ferreira (2003) point out that the idea of "filtering" out disfluencies (despite their prosody remaining intact, location not easily remembered by listeners) does not account for the *incremental* nature of speech processing; i.e. the processing starts before the input is complete. Filtering would imply that processing needs to occur *after* the removal of disfluencies, and that the listener then, would need to wait for the entire utterance to be completed by the speaker before processing. However, the processes involved in comprehension are continuous and *incremental*, as humans process utterances incrementally. Listeners *must* have developed a comprehension system in order to process disfluencies, given how frequently they occur in spoken language (Bailey and Ferreira, 2003). Disfluencies are part of the incremental processing of the

6. Confusions in terminology arise from research itself, as "disfluencies" is an all-encompassing term used for many works, including when the works are *only* concerned with fillers.

flow of speech; for instance Bailey and Ferreira (2003) show that disfluencies can affect the internal syntactic parser of the listener. The disfluencies considered here was the filler "uh", but also any kind of interruption (noise) was also deemed to be a disfluency. Brennan and Schober (2001) show that listeners may use disfluencies as cues to avoid integrating what they deem to be incorrect material in an online processing task. Thus, there is evidence to support that fillers are not filtered out by the listener, and indeed, included in the *online processing of the utterance*.

Fillers used as cues to understand new information has even been shown *neurologically*, and in an online processing task. Corley *et al.* (2007) studied the effect of filler ("um", called a "hesitation" in the work) on the listener's comprehension using the $N400$ function of an Event-related potential (ERP). The $N400$ effect can be observed during language comprehension, typically occurring 400 ms after the word onset; it is a negative charge recorded at the scalp consequent to hearing an unpredictable word. The $N400$ effect was first established in listeners who heard unpredictable words compared to predictable words. Then, when the filler proceeded the unpredictable word, the $N400$ effect in listeners was *visibly reduced*. In a subsequent memory test on the listener, words preceded by this filler were more likely to be remembered.

2.2.4. *Time-buying measures and effects on memory*

Works also hypothesised whether all these effects of integration can be explained by the *processing time hypothesis*. This means, considering whether the disfluent speech is more memorable/noticeable simply because disfluencies add more time to the speech utterance (referring to research that studies the role of fillers in giving pause to the discourse). Thus, does this effect – i.e. a pause in the utterance – cause the listener to simply give more attention to the utterance? Fraundorf and Watson (2011) examined this in a study on how fillers affect the memory of the listeners. They exposed the listeners to fluent speech containing fillers versus coughs of equal duration artificially spliced into the speech. They found that while fillers facilitated recall, therefore being beneficial on memory, coughs negatively hampered recall accuracy. Disfluent speech (fillers) is hence more likely to be remembered by the listener, and this is *not solely based on the additional time added to the utterance*. Fraundorf and Watson (2011) also manipulated the location of the fillers in speech to study the effect of *position* of fillers on comprehension. This was based on the findings of Swerts (1998) (discussed previously), who found that following fillers, listeners may expect a speaker to shift topics as they carry information about larger topical units – therefore, acting as cues for discourse structure. However, Fraundorf and Watson (2011) found that fillers benefit listener's recall accuracy regardless of its typical or atypical location. Tottie (2014) found that fillers are noticed more when overused or used in the wrong context, so while they may facilitate recall regardless of location, they still may be more noticeable in atypical locations.

Thus the speaker produces disfluencies, and we have illustrated works that show that they may be strategically used, but also that they can indicate problems in the planning process. When comparing production versus comprehension, regardless of

whether the speaker intentionally uses disfluencies (Barr and Seyfeddinipur, 2010), the listener still *integrates* them as cues for upcoming information. However, it seems that the type of disfluency may also matter in the process of comprehension and whether the listener integrates these disfluencies or filters them out.

3. Computational Perspectives: Removal versus Integration

We briefly defined the computational perspective on disfluencies; i.e. works concerned with the computational processing of disfluencies in order to remove them; rendering the utterance more "fluent" and text like. In this distinction, we also consider linguistic work motivated by this goal. For example, phonetic works that characterise the acoustic environments that disfluencies occur in, with the ultimate goal of aiding Automatic Speech Recognition (ASR) systems. While it is a general trend to consider *all* disfluencies as noise in this perspective, we also distinguish research that has found disfluencies informative in SLU. SLU is broad and not one defined task, it "combines speech processing and natural language processing (NLP) by leveraging technologies from machine learning (ML) and artificial intelligence (AI)" (Tur and De Mori, 2011). Thus we distinguish between what is generally considered SLU (SLU for Spoken Dialogue Systems), and broader SLU (SLU tasks concerned with social communication), loosely adopting the distinction made in Tur and De Mori (2011). We make this distinction, because the treatment of disfluencies varies greatly – from *removal to integration* – depending on the SLU task.

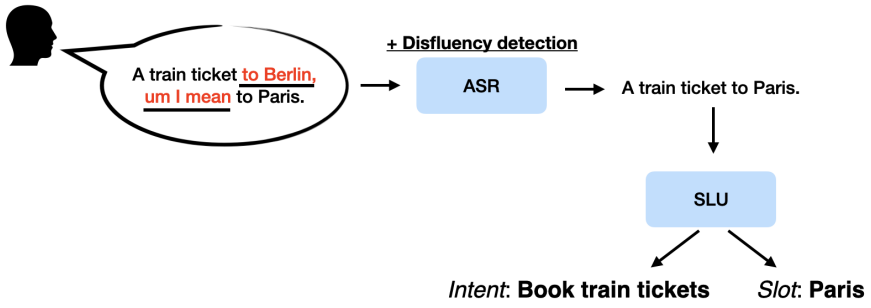


Figure 1. *Example of SLU for SDS, where the disfluent part of the utterance is highlighted in red and underlined. As shown, the disfluencies are removed in post-processing using disfluency detection systems, after ASR systems transcribe the input speech. Then, the input is further collapsed into a semantic frame, consisting of intent and slot.*

3.1. *SLU for Spoken Dialogue Systems (SDS)*

In SLU for SDS, the objective is to collapse the input utterance into a *semantic frame*, consisting of an *intent* and a *slot* (Louvan and Magnini, 2020).

Consider the disfluent utterance "A train ticket to Berlin uh I mean to Paris". If "to Berlin uh I mean" is not removed (marked in red in Fig. 1 to get the sanitised utterance; "A train ticket to Paris"), it leads to confusion in the subsequent (semantic) processing of the utterance – the correct intent would be identified (i.e. book train tickets) and but not the correct slot (i.e "Paris" not "Berlin"). Disfluencies in this regard are discarded as noise. Thus in a standard SDS pipeline, the output transcripts of ASR systems are cleaned of disfluencies in post-processing as shown in Fig. 1, using disfluency detection systems. From a computational perspective, the intent to process of disfluencies in present day is due to automatic Natural Language Understanding (NLU) systems not being *robust* to them (Ginzburg *et al.*, 2014). Representations of this semantic frame focus on the lexical aspect of the utterance but neglect the non-lexical channel. Thus removing disfluencies such as fillers could remove important information about the *social* aspects of communication (subsequently discussed in Sec. 3.2).

There has been an interest in studying the characteristics of disfluencies for recognition purposes, such as how they are distributed in corpora (for instance, Shriberg (1994) and Shriberg (2001)), and also, at different linguistic levels. The motivation of these works was to identify the properties of disfluent speech to *enable a comparison with its fluent counterparts*. For instance, at a phonetic level, Shriberg (1999) found that the vowels in fillers gave much longer durations than the same vowels in fluent contexts, and Shriberg and Lickley (1993) found that the intonation of fillers is not independent of prior prosodic context. These studies were done with the aim of ideally *leveraging these features for ASR to be more robust to disfluencies*, but with mixed results (such as in Shriberg *et al.* (1997)). Other linguistic levels have also been studied, such as morpho-syntactic features (Goryainova *et al.*, 2014), contextual occurrence (Vasilescu *et al.*, 2010), pragmatic levels (Shriberg *et al.*, 1998) and so on.

3.1.1. *Disfluency detection*

Johnson and Charniak (2004) were the first to introduce the task of disfluency detection, framing it as a noisy channel problem; i.e. noise (disfluencies) has been added to the fluent source utterance. Presently, the task of detection is framed in many ways. For instance, as a *sequence-tagging* task, where typically each token in the input utterance needs to be classified as either in the beginning, inside, or outside (BIO) of a disfluent region. This has been adapted to neural architectures, for instance Zayats and Ostendorf (2019), with a bidirectional Long Short Term Memory (biLSTM) network, or a Transformer model from Lou and Johnson (2020b). Encoder-decoder frameworks allow for detection to be framed as a *sequence-to-sequence neural translation task*, where the encoder learns a representation of the disfluent utterance, and the decoder generates a fluent version of that utterance (see Wang *et al.* (2018)). Similar to this, the objective of end-to-end *speech translation* is to output fluent text from an input of disfluent *audio*. This frames the detection problem as an interim step between ASR and a downstream task such as machine translation (see Salesky *et al.* (2019)). Lou and Johnson (2020a) propose an end-to-end speech recognition and disfluency removal system, so that the fluent output may be used for any downstream

task. Hough and Schlangen (2017) propose a *joint task* of disfluency detection and utterance segmentation. They find that the task is better approached jointly, with the combined system outperforming results on individual tasks. Detection may be combined with other tasks in a multi-task learning (MTL) scenario, showing improvements when there is a secondary task of language modelling (Shalyminov *et al.*, 2018), or even simultaneous tasks of detection, POS-tagging, language modelling and utterance segmentation (Rohanian and Hough, 2020).

A drawback on state-of-the-art (SOTA) systems is they are not robust to transcribing disfluencies correctly in-the-wild, with criticisms that they could not work in an online processing task (Chen *et al.*, 2022) nor with raw transcripts from ASR (Rohanian and Hough, 2021). Specifically for *disfluency detection*, often the *textual input* used is conditioned on already appropriate segmentation and accurate transcription – which is unrealistic in a real-world scenario. From Rohanian and Hough (2021), works on disfluency detection "are almost *exclusively* conducted on pre-segmented utterances of the Switchboard (SWBD) (Godfrey *et al.*, 1992) corpus of telephone conversations"; contradicting the generalisation capabilities of such systems. Works are exploring the feasibility of detection with online processing (Chen *et al.*, 2022; Rohanian and Hough, 2021; Shalyminov *et al.*, 2018; Hough and Schlangen, 2017), with transcripts arising from ASR (Rohanian and Hough, 2021). With hybrid conversations, where user interactions switch between task-oriented and open domain requests (Kim *et al.*, 2021), this topic of detecting disfluencies is starting to come to the forefront once more. The improvement of these systems would be beneficial for other types of research, i.e. better automatic annotation of disfluencies in corpora in order to study disfluent phenomena – particularly for areas such as clinical SLU, to help determine a patient's cognitive state.

3.1.2. *Filler detection*

Systems may be better adapted to recognising specific kinds of disfluencies depending on their properties; fillers for instance, are acoustically distinct. Off-the-shelf open-source speech recognisers such as CMU Sphinx (Lamere *et al.*, 2003) provide functionality to define a filler phone dictionary of *paralinguistic sounds* that are not present in the Language Model (LM). Das *et al.* (2019) leverage the acoustic properties of fillers by training a convolutional recurrent neural network (CRNN) applied directly to *audio recordings* to do filler segmentation, i.e. segmenting fillers in order to output a more fluent utterance. Fillers here are considered to be "unprofessional speech", stating that the speaker sounds more "fluent, confident and practised compared to the original recorded speech" evaluated by automatic measures [7], rather than human annotation judgements. Regardless, it would be beneficial to many branches of research to automatically label fillers reliably. However, as pointed out in Das *et al.* (2019), these systems can suffer from false positives (for instance the "um" in

7. Proposed in Kormos and Dénes (2004) to study the perception of fluency in second language learners. Indeed they state in the work that the number of fillers *did not affect the perception of fluency the raters had* (see further discussion in Sec. 3.2, Broader SLU).

"umbrella"). Recently, a new filler detection benchmark and dataset was introduced, where the objective was to label filler candidates given an incoming speech input (Zhu *et al.*, 2022). The system leverages voice activity detection (VAD) and ASR to detect possible filler candidates from other tokens, and then a classifier to categorise them. They evaluate the ASR approach to find it outperforms a keyword spotting approach. The publicly available dataset is based on podcast episodes and contains a diverse range of speaking styles and topics – which could facilitate many future research directions. Thus, there is attention now given to individually detecting fillers from raw audio as a task, without gold-standard transcripts.

Interestingly, while Google Cloud Speech-to-Text services map paralinguistic sounds such as fillers to silence, it has introduced a functionality in its medical ASR model [8] to include fillers when transcribing medical documents. To contrast, in the Google Dialogflow API [9], there are guidelines given in order to *avoid* fillers as giving unnecessary variety to utterances, as they are ignored by subsequent NLU models (that collapse the input into a semantic frame as shown in Fig. 1). What is interesting here is the *exclusion* of fillers in an SLU for SDS task (i.e. Google Dialogflow), and the *inclusion* of fillers in an ASR task with a broader SLU context – i.e. to pick up social cues in medical transcription. Thus the treatment of fillers and disfluencies is evolving in SLU research to consider specifically the nature of the task; compared to the previous view that all disfluencies must be removed as noise. Hence detection may focus on *removal* if the subsequent task is to collapse the input into a semantic frame, or *integration* if the social aspects of communication may be further considered.

3.1.3. *Annotation for disfluency detection*

Annotation schemes for disfluency detection are almost always form based, as the intent is to sanitise the utterance of disfluencies. For this, there needs to be precise, form-based characterisations of the environment in which disfluencies occur, with no interest in deliberating as to why there was a departure from fluency. We briefly outline the most commonly used annotation scheme in disfluency detection, consistent with the Switchboard repair mark-up (Meteer *et al.*, 1995). Consider the following example:

$$\text{Archie } \underbrace{[\text{ likes}}_{\text{RM}} + \underbrace{\{\text{F uh }\}}_{\text{IM}} \underbrace{\text{ loves }]}_{\text{RP}} \text{ Veronica.} \qquad [1]$$

At a high level, there is the notion of *some kind of erroneous speech that is to be replaced by corrected speech* – here,"loves" to replace "likes". This is used as a starting point for annotation (originally proposed by Levelt (1983)). While variations of this scheme exist, the underlying reasoning *specifically* for detection tasks remains the same. Formally, there is: i) the *reparandum phase* (RM), i.e. or the entire region to be

8. https://cloud.google.com/speech-to-text/docs/medical-models.
9. https://cloud.google.com/dialogflow/cx/docs/concept/agent-design.

deleted, and ii) the *repair phase* (RP), i.e. what replaces the RM. This was adopted by Shriberg (1994), who also proposes the term iii) *interregnum phase* IM, which is an optional interruption point, where the speaker may realise that a correction needs to be made. *Non-sentence elements* (such as fillers) can occur within this IM structure or outside, also called *isolated edit terms/ single edit tokens*. From this structure, several types of disfluencies can be formally defined. For instance *repetitions* – i.e. when the RM phase is repeated exactly in the RP phase, and *restarts* – i.e. when the RM phase is discarded. This annotation structure is useful for disfluency detection and SLU because the reasoning is to only keep the corrected speech in the utterance (i.e. the RP) for downstream processing (see Fig. 1), and discard the rest as noise (RM+IM).

However, a drawback of this scheme for a detection task is that it spans only one speaker turn and can only be initiated by the speaker producing the disfluencies. In many contexts, repair could span different speaker turns and different interlocutors (see a detailed discussion in Purver *et al.* (2018)). Disfluency detection thus can fail on longer disfluencies (Zayats *et al.*, 2019), and disfluencies spanning multiple turns (Purver *et al.*, 2018). Additionally, many other annotation schemes exist when the objective is not constrained to detection alone. For instance, Christodoulides *et al.* (2014) proposed a tool to holistically do a form-based annotation of several phenomena characteristic of spontaneous speech: with disfluency detection and annotation, and multi-word unit recognition (including POS, syllable, phone tagging ...). Rather than an annotation specific to one task, these layers of annotation could be beneficial for several downstream SLU tasks. The annotation scheme is more realistically designed for spoken language – such as using prosodic cues for segmentation in the absence of punctuation. There are also language specific schemes proposed (such as Benzitoun *et al.* (2012), Kahane and Gerdes (2020) and Eshkol *et al.* (2010) for French), as **a limitation of this survey is that we constrain ourselves to English** [10].

Problems may also arise when transcribing (disfluent) audio itself. Le Grezause (2017) found that transcribers who had transcribed a few conversations tended to have *substantially higher transcription errors*, compared to transcribers that had transcribed a large number of conversations. Zayats *et al.* (2019) found that tokens generally related to spontaneous speech phenomena (fillers ...) have a high frequency of transcription errors, which they *discuss could be due to these tokens being non-standard*, i.e. unaccounted for in annotation instructions. Fillers in particular, are among some of the most likely tokens to be mis-transcribed (whether inserted, deleted or substituted), and transcriber experience has a noticeable effect on the accuracy of transcribing fillers specifically (Le Grezause, 2017).

10. Note also that the issue of annotation has been extensively studied from a non-computational linguistic perspective (Grosman, 2018), with extensive summaries of terminology from different linguistic works (Lickley *et al.*, 1991; Nicholson, 2007; Lickley, 2015), and new and emergent annotation schemes proposed (Crible *et al.*, 2015) which are not discussed here.

3.1.4. *Text representations of disfluencies*

With the increasing popularity of voice assistant technologies, a trend has emerged in Natural Language Understanding (NLU) research to overlap with SLU; i.e. considering the *textual processing of speech transcripts* (as discussed in Ruder (2020) as "Speech first-NLP"). However, discrepancies may arise in the processing of speech transcripts compared to processing grammatically written text, if utilising the same (NLP) systems. For example, Tran *et al.* (2017) present an attention-based encoder-decoder model for parsing conversational sentences arising from speech transcripts. An important finding from this work was that the integration of acoustic-prosodic features showed *the most gains over disfluent and longer sentences* compared to fluent ones. This empirically *shows a discrepancy* between transcripts that are more "speech-like" (i.e. are disfluent and have acoustic variation) compared to transcripts that are more grammatical and resemble written text. It is worth noting that for this task of parsing, there is the assumption of *already structured* spontaneous speech transcripts to be used as input (i.e. annotated punctuation, input being pre-segmented sentences ...) – so indeed, there is a further gap between perfectly fluent, grammatical text versus raw ASR/verbatim transcripts. Hervé *et al.* (2022) investigate how deep contextualised embeddings could be pre-trained on massive amounts of ASR generated transcripts for more realistic samples in spoken language modelling. The downstream tasks including parsing conversational utterances show improvements of the model pre-trained on raw ASR data (which they call FlauBERT-Oral) compared to the original model trained on written data (i.e. FlauBERT, a French LM (Le *et al.*, 2020)). They point out that the ASR generated only lowercase transcripts, and that adding capitalisation and punctuation to the transcripts could benefit the model i.e. *re-introducing some degree of sentence structure*.

Tran *et al.* (2019) showed that deep contextualised embeddings pre-trained on large written corpora can be fine-tuned on smaller spontaneous speech datasets to improve parsing on conversational speech transcripts. While this indicates that some *general* characteristics of spontaneous speech may be learnt in the fine-tuning stage (and as shown by Hervé *et al.* (2022) using models pre-trained on raw ASR), results are mixed when specifically considering the text representation of fillers. Barriere *et al.* (2017) for instance showed that pre-trained word embeddings such as Word2vec (Mikolov *et al.*, 2013) have poor representation of spontaneous speech phenomena such as "uh", as they are trained on written text and do not carry the same meaning as when used in speech. Interestingly, this discrepancy may not always negatively impact the *robustness* of the system when purely considering fillers. For instance, Dinkar *et al.* (2020) investigated the representations of fillers using deep contextualised word embeddings. They found that Bi-directional Encoder Representations (BERT) (Devlin *et al.*, 2019) has existing representations of fillers, despite being pre-trained on massive amounts of written text. They found (similar to Stolcke and Shriberg (1996)'s results on $ngrams$) that the inclusion of fillers *reduces* the uncertainty of a language model in a spoken language modelling task. This is despite research that shows that overall, speech disfluencies occur at higher perplexities (Sen, 2020). Thus, in addition to psycholinguistic work, there is computational research to show that indeed, fillers

may be able to provide information regarding the neighbouring words to the right in a language modelling task. However, Dinkar *et al.* (2020) found that BERT is unable to distinguish between the two fillers "uh" and "um", despite research to show that they occupy different functions in discourse (Le Grezause, 2017; Dinkar, 2022). To conclude, there is now the general awareness that spoken speech is not like written text, and increasing studies work on how to learn representations of spontaneous speech given that architectures are usually developed on large amounts of written data.

3.2. *Broader SLU and Generation*

3.2.1. *Broader Spoken Language Understanding (SLU)*

We consider broader SLU as the analysis of the flow of speech by leveraging NLP and ML techniques [11] – particularly for higher order tasks that are concerned with social/affective computing. Though the methodologies differ from the psycholinguistic approach to study the production contexts of disfluencies, the underlying intent is similar. That is, to study the context of disfluencies produced (and indeed, many other lexical and non-lexical features) and their link to a variety of socio-communicative and cognitive phenomena. The findings are useful for feature engineering in broader SLU tasks. For example, disfluencies in several works have been found to be an informative *social signal* (Ekman *et al.*, 1980; Mairesse *et al.*, 2007; Vinciarelli and Mohammadi, 2014; Schuller *et al.*, 2019). Fillers specifically are commonly used as an attribute to study big 5 personality traits (Mairesse *et al.*, 2007). The Computational Paralinguistics Challenge focused on detecting fillers, which they considered to be a "social signal" (Schuller *et al.*, 2019), acknowledging the importance of fillers in *tasks concerned with social communication*. Along these lines, research in personality computing has the most consistent correlation, with observations made from speech; including *paralanguage* such as fillers (Ekman *et al.*, 1980; Vinciarelli and Mohammadi, 2014). We cannot exhaustively account for these works, as they are numerous. Some examples are research on the role of disfluencies (mainly fillers) in the prediction of a speaker's *emotions* (Moore *et al.*, 2014; Tian *et al.*, 2015), *stance* (Le Grezause, 2017), perceived *persuasiveness* (Park *et al.*, 2014), perceived *confidence* (Dinkar *et al.*, 2020), etc. Dufour *et al.* (2014) point out that spontaneous speech is not the same as prepared speech, where the utterances are well formed and closer to written documents. They focus on classifying the *degree of spontaneity* of speech in order to do SLU tasks, such as characterising speaker roles (example, an interviewer versus an interviewee) using fillers and other acoustic-linguistic features. Recent work argues that disfluencies will be useful in dialogue based computer-assisted language learning; i.e. detecting and analysing a learner's disfluencies (including silences and laughter) could potentially help a system determine appropriate pedagogical interventions (Skidmore and Moore, 2022). Disfluencies may play a crucial role in *clinical SLU* tasks, such as in dementia recognition (Rohanian *et al.*, 2021). As stated, Google

11. Which we loosely adopt from Tur and De Mori (2011).

Cloud Speech-to-Text has included in its medical ASR model options to include fillers, for the purpose of transcribing medical documents.

However, an open challenge that remains in analysis is the inclusion of context. Research may predetermine based on the scenario whether disfluencies are positive or negative. For instance, in the automatic processing of job interview data (Rasipuram *et al.*, 2016), fluency is assumed to be desirable. While indeed, the speaker's production of disfluencies may have an effect on the outcome of a job interview; there are nuances of how the speaker utilises such spontaneous speech phenomena; i.e. both production and perception will vary based on socio-linguistic background, context, domain and so on. Analysis may be based on the idea that the frequency of words (including, non-lexical tokens such as fillers present in transcripts) can represent underlying affective traits (Boyd and Schwartz, 2021). In this aspect, the link between fillers and a wide variety of phenomena is to be expected – from linguistic levels to higher affective levels; with Barr (2001) even describing fillers as *vocal gestures*. In a recent survey, Boyd and Schwartz (2021) discuss this drawback regarding research in the intersection between psychology and language analysis – but also considering interdisciplinary fields such as social computing; stating that it may not always be a case of "paying attention to X is correlated with Y". Thus we should cautiously assume a linear relationship between the fillers produced and the task under consideration. Dinkar *et al.* (2021) for instance found that speakers tend to stylistically use fillers before introducing new information in the discourse, but that listeners may not associate this specific use of fillers with their estimation of the speaker's confidence.

An additional challenge is the varied terminology used to describe fillers in the context of broader SLU. While standardised terminology may be used in a benchmark task such as detection, it is not the case for less standardised tasks. This is problematic, because it can lead to a lack of availability and transparency of the findings. For instance, the *Linguistic Inquiry and Word Count (LIWC)* (Pennebaker *et al.*, 2001), a commonly used text analysis software in personality computing (Mehta *et al.*, 2020), gives guidelines to annotate "nonfluencies" (fillers) – "uh", "um", "er", and "stuttering". Here, "stuttering" broadly denotes the general phenomena of being disfluent [12].

3.2.2. *Generation*

Similar to psycholinguistic approaches, it is to be noted that there are works that study the *perception* of disfluencies from a generation standpoint, i.e. using artificially synthesised disfluencies in speech. While this is not under the umbrella of SLU as such, we briefly describe some research specific to fillers. For instance, in a work directly motivated by psycholinguistic perspectives of comprehension, Wollermann *et al.* (2013) explore the listener's perception of disfluencies using Text-to-Speech (TTS). This was based on the work of Brennan and Williams (1995), that discusses

12. Please note, the term is not to be confused with the clinical sense of "stuttering". In clinical literature, the term commonly used is "stutter-like disfluencies" (SLDs). This use seems borrowed from Mahl (1956), who used the term to refer to repetition of partial words (Lickley, 2015).

the role of fillers and prosodic cues in a listener's evaluation of how uncertain they think the speaker is regarding a topic. They had the system exhibit "uncertain" behaviour through disfluent TTS responses in a question-answering context. They found that disfluencies in combination with prosodic cues (i.e. delays + fillers) increased a listener's perception of uncertainty towards the system's answers. Similarly, Kirkland *et al.* (2022) found that when fillers were not present in synthesised speech, it leads to a perception of more confident sounding utterances, while utterance-medial fillers lead to a perception of the least-confident sounding utterances (both in addition to other prosodic features).

Some works may not have basis in psycholinguistic theory. For instance, there is research that considers how disfluencies enhance the *naturalness* of the synthesised speech. Pfeifer and Bickmore (2009) evaluate an agent that uses fillers "uh" and "um" in speech. The motivation behind this was to improve the naturalness of speech in an Embodied Conversational Agent (ECA), as ECAs often try to emulate humans in gestures and facial expressions, yet speak in fluent sentences. Results are mixed, with some participants saying that fillers enhanced the naturalness of the conversation, while others expected that an agent should speak fluently, and fillers were deemed inappropriate. Székely *et al.* (2019) discuss approaches for treating fillers in TTS tasks, i.e. suggesting methods that will result in them being synthesised naturally (both distributionally and perceptually) in the generated output. Disfluencies may also be utilised as a *communicative strategy* in generation. Skantze *et al.* (2015) studied how a system can use multi-modal turn-taking signals (including fillers) as a *time-buying measure*, i.e. to buy time for generating a response as the next move of the robot is decided.

While a common goal of AI is to work towards more human-like (anthropomorphic) agents, a challenge that remains is to consider the trade-off between the naturalness of a system and the safety of its deployment. Consider Google Duplex (Leviathan and Matias, 2018), a TTS system for accomplishing real-world tasks over the phone. The *inclusion of disfluencies* (such as fillers and repairs) led to highly natural sounding generated responses, showing how ubiquitous disfluencies are to everyday communication. However, these responses made human callers think that they were conversing with another human. This illusion of agency may have negative consequences when considering *safety* in conversational AI. For example, attributing anthropomorphic traits to an agent in an IMPOSTOR EFFECT (Dinan *et al.*, 2021; Abercrombie and Rieser, 2022) scenario, i.e. where a system provides inappropriate advice in safety critical situations (such as when a user seeks medical advice). Despite this, from a generation perspective, fillers and disfluencies may still be desirable if used as a mitigation strategy, i.e. to exhibit uncertainty and doubt in safety critical situations (e.g. "uh… I'm not an expert here but…"). This is important, as Mielke *et al.* (2022) pointed out that neural dialogue agents are not linguistically calibrated – i.e. despite the agent being *factually inaccurate* it may still (inappropriately) verbalise expressions of confidence in responses. Thus, *the trade-off between naturalness and safety* merits consideration in future research, particularly when using disfluencies in generation tasks.

4. Conclusion

The preliminary goal of this article was to bring together several different perspectives on fillers that influence research in SLU. The article was motivated by the increasing popularity of voice assistant technologies, where often, the corpora used will invariably contain these spontaneous speech events. Yet, to the best of our knowledge, there was not an introductory source on these specific phenomena, that offered a *breadth* of different perspectives and approaches – particularly targeted towards the SLU community. We were motivated to introduce these different perspectives, without exhaustively going into details that may overwhelm the reader. Our aim was to discuss these perspectives in a holistic way, that is considering them at various stages of the SLU pipeline; from underlying (pyscho)linguistic theory, to annotation, ASR and SLU perspectives, and finally, looking briefly at research on the generation of these phenomena. We have tried to the best of our ability to synthesise the main perspectives in an approachable way, and suggest further reading and other sources when appropriate. Additionally in each section we pinpoint (what we believe) are the main challenges and trends of each area. To do so, we identified two areas of research, i.e. psycholinguistic and computational perspectives. In the former, we focused on the research available on the production and comprehension of disfluencies, particularly fillers. In the computational approaches, we discussed the treatment of disfluencies by distinguishing between SLU as it is typically considered for SDS, and broader SLU tasks more concerned with social communication. We try to highlight the commonalities regarding both perspectives as well, i.e. discussing works in both areas that find disfluencies informative, consider disfluencies as noise, consider the listener's perception .... Going forward, some open challenges are to consider the balance between removing disfluencies with regard to the robustness of a system versus including them as speech events that could offer social context, the trade-off between how natural sounding the system is versus how safe it may be in deployment, the integration of fine-grained context, and lastly, the taxonomy of such phenomena in order to make the research on them more accessible.

Acknowledgements

Tanvi Dinkar is supported by 'AISEC: AI Secure and Explainable by Construction' (EP/T026952/1), and Tanvi Dinkar and Chloé Clavel were supported by 'ANIMATAS: Advancing intuitive human-machine interaction with human-like social capabilities for education in schools', European Union's Horizon 2020 research and innovation programme under grant agreement No. 765955.

5. References

Abercrombie G., Rieser V., "Risk-graded Safety for Handling Medical Queries in Conversational AI", *arXiv preprint arXiv:2210.00572*, 2022.

Arnold J. E., Kam C. L. H., Tanenhaus M. K., "If you say thee uh you are describing something hard: the on-line attribution of disfluency during reference comprehension.", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 33, n° 5, p. 914, 2007.

Arnold J. E., Tanenhaus M. K., Altmann R. J., Fagnano M., "The old and thee, uh, new: Disfluency and reference resolution", *Psychological science*, vol. 15, n° 9, p. 578-582, 2004.

Bailey K. G., Ferreira F., "Disfluencies affect the parsing of garden-path sentences", *Journal of Memory and Language*, vol. 49, n° 2, p. 183-200, 2003.

Bard E. G., Lickley R. J., Aylett M. P., "Is disfluency just difficulty?", *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech*, 2001.

Barr D. J., "Trouble in mind: Paralinguistic indices of effort and uncertainty in communication", *Oralité et gestualité: Interactions et comportements multimodaux dans la communication*p. 597-600, 2001.

Barr D. J., Seyfeddinipur M., "The role of fillers in listener attributions for speaker disfluency", *Language and Cognitive Processes*, vol. 25, n° 4, p. 441-455, 2010.

Barriere V., Clavel C., Essid S., "Opinion Dynamics Modeling for Movie Review Transcripts Classification with Hidden Conditional Random Fields", *Proceedings of Interspeech 2017*, Stockholm, Sweden, August, 2017.

Beattie G. W., Butterworth B. L., "Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech", *Language and speech*, vol. 22, n° 3, p. 201-211, 1979.

Bell A., Jurafsky D., Fosler-Lussier E., Girand C., Gregory M., Gildea D., "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation", *The Journal of the Acoustical Society of America*, vol. 113, n° 2, p. 1001-1024, 2003.

Bender E. M., Koller A., "Climbing towards NLU: On meaning, form, and understanding in the age of data", *Proc. of ACL*, 2020.

Benzitoun C., Fort K., Sagot B., "TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe (TCOF-POS : A Freely Available POS-Tagged Corpus of Spoken French) [in French]", *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, ATALA/AFCP, Grenoble, France, p. 99-112, June, 2012.

Boyd R. L., Schwartz H. A., "Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field", *Journal of Language and Social Psychology*, vol. 40, n° 1, p. 21-41, 2021.

Brennan S. E., Schober M. F., "How listeners compensate for disfluencies in spontaneous speech", *Journal of Memory and Language*, vol. 44, n° 2, p. 274-296, 2001.

Brennan S. E., Williams M., "The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers", *Journal of memory and language*, vol. 34, n° 3, p. 383-398, 1995.

Chen A., Zayats V., Walker D. D., Padfield D., "Teaching BERT to Wait: Balancing Accuracy and Latency for Streaming Disfluency Detection", 2022.

Christodoulides G., Avanzi M., Goldman J.-P., "DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. An Evaluation on a Corpus of French Spontaneous and Read Speech", *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), Reykjavik, Iceland, May, 2014.

Clark H. H., Fox Tree J. E., "Using uh and um in spontaneous speaking", *Cognition*, vol. 84, n° 1, p. 73-111, 2002.

Colman M., Healey P., "The distribution of repair in dialogue", *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 33, 2011.

Corley M., MacGregor L. J., Donaldson D. I., "It's the way that you, er, say it: Hesitations in speech affect language comprehension", *Cognition*, vol. 105, n° 3, p. 658-668, 2007.

Corley M., Stewart O. W., "Hesitation disfluencies in spontaneous speech: The meaning of um", *Language and Linguistics Compass*, vol. 2, n° 4, p. 589-602, 2008.

Crible L., Dumont A., Grosman I., Notarrigo I., "Annotation des marqueurs de fluence et disfluence dans des corpus multilingues et multimodaux, natifs et non natifs v. 1.0", 2015.

Das S., Gandhi N., Naik T., Shilkrot R., "Increase Apparent Public Speaking Fluency by Speech Augmentation", *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 6890-6894, 2019.

Devlin J., Chang M.-W., Lee K., Toutanova K., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, p. 4171-4186, June, 2019.

Dinan E., Abercrombie G., Bergman A. S., Spruit S., Hovy D., Boureau Y.-L., Rieser V., "Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling", 2021.

Dinkar T., Computational models of disfluencies: fillers and discourse markers in spoken language understanding, PhD thesis, Institut Polytechnique de Paris, 2022.

Dinkar T., Colombo P., Labeau M., Clavel C., "The importance of fillers for text representations of speech transcripts", *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, p. 7985-7993, November, 2020.

Dinkar T., LTCI T. P., Biancardi B., Clavel C., "From local hesitations to global impressions of a speaker's feeling of knowing", *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue*, 2021.

Dufour R., Esteve Y., Deléglise P., "Characterizing and detecting spontaneous speech: Application to speaker role recognition", *Speech communication*, vol. 56, p. 1-18, 2014.

Ekman P., Friesen W. V., O'Sullivan M., Scherer K., "Relative importance of face, body, and speech in judgments of personality and affect", *Journal of personality and social psychology*, vol. 38, n° 2, p. 270, 1980.

Eshkol I., Maurel D., Friburger N., "Eslo: From Transcription to Speakers' Personal Information Annotation", *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA), Valletta, Malta, May, 2010.

Fox Tree J. E., "The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech", *Journal of memory and language*, vol. 34, n° 6, p. 709-738, 1995.

Fraundorf S. H., Arnold J., Langlois V. J., "Disfluency", 2018.

Fraundorf S. H., Watson D. G., "The disfluent discourse: Effects of filled pauses on recall", *Journal of memory and language*, vol. 65, n° 2, p. 161-175, 2011.

Ginzburg J., Paris-diderot U., Fernández R., Schlangen D., "Disfluencies as intra-utterance dialogue moves", *Semantics and Pragmatics*, 2014.

Godfrey J. J., Holliman E. C., McDaniel J., "SWITCHBOARD: Telephone Speech Corpus for Research and Development", *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, IEEE, p. 517-520, 1992.

Goryainova M., Grouin C., Rosset S., Vasilescu I., "Morpho-Syntactic Study of Errors from Speech Recognition System", *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), Reykjavik, Iceland, p. 3045-3049, May, 2014.

Grosman I., Évaluation contextuelle de la (dis)fluence en production et perception: pratiques communicatives et formes prosodico-syntaxiques en français, PhD thesis, UCL-Université Catholique de Louvain, 2018.

Hervé N., Pelloin V., Favre B., Dary F., Laurent A., Meignier S., Besacier L., "Using ASR-Generated Text for Spoken Language Modeling", *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, Association for Computational Linguistics, virtual+Dublin, p. 17-25, May, 2022.

Hough J., Schlangen D., "Joint, incremental disfluency detection and utterance segmentation from speech", *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, p. 326-336, 2017.

Johnson M., Charniak E., "A TAG-based noisy-channel model of speech repairs", *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, p. 33-39, 2004.

Kahane S., Gerdes K., "Annotation syntaxique du français parlé: Les choix d'ORFÉO", *Langages*, n° 3, p. 69-86, 2020.

Kim S., Liu Y., Jin D., Papangelis A., Gopalakrishnan K., Hedayatnia B., Hakkani-Tur D., "How Robust RU?: Evaluating Task-Oriented Dialogue Systems on Spoken Conversations", *arXiv e-prints*p. arXiv-2109, 2021.

Kirkland A., Lameris H., Székely E., Gustafson J., "Where's the uh, hesitation? The interplay between filled pause location, speech rate and fundamental frequency in perception of confidence", *Proc. Interspeech 2022*p. 4990-4994, 2022.

Kormos J., Dénes M., "Exploring measures and perceptions of fluency in the speech of second language learners", *System*, vol. 32, n° 2, p. 145-164, 2004.

Lamere P., Kwok P., Gouvea E., Raj B., Singh R., Walker W., Warmuth M., Wolf P., "The CMU SPHINX-4 speech recognition system", *Ieee intl. conf. on acoustics, speech and signal processing (icassp 2003), hong kong*, vol. 1, p. 2-5, 2003.

Le Grezause E., Um and Uh, and the expression of stance in conversational speech, PhD thesis, 2017.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A., Crabbé B., Besacier L., Schwab D., "FlauBERT: Unsupervised Language Model Pre-training for French", *Proceedings of the Twelfth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 2479-2490, May, 2020.

Levelt W. J., "Monitoring and self-repair in speech", *Cognition*, vol. 14, n° 1, p. 41-104, 1983.

Leviathan Y., Matias Y., "Google Duplex: An AI System for Accomplishing Real World Tasks Over the Phone.", *Google AI Blog*, 2018.

Lickley R. J., Detecting disfluency in spontaneous speech, PhD thesis, University of Edinburgh, 1994.

Lickley R. J., "Fluency and Disfluency", *The handbook of speech production*p. 445, 2015.

Lickley R. J., Bard E. G., "When can listeners detect disfluency in spontaneous speech?", *Language and speech*, vol. 41, nº 2, p. 203-226, 1998.

Lickley R. J., Shillcock R. C., Bard E. G., "Processing Disfluent Speech: How and when are disfluencies found?", *Second European Conference on Speech Communication and Technology*, 1991.

Lou P. J., Johnson M., "End-to-End Speech Recognition and Disfluency Removal", *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, Online, p. 2051-2061, November, 2020a.

Lou P. J., Johnson M., "Improving Disfluency Detection by Self-Training a Self-Attentive Model", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, p. 3754-3763, July, 2020b.

Louvan S., Magnini B., "Recent Neural Methods on Slot Filling and Intent Classification for Task-Oriented Dialogue Systems: A Survey", *arXiv preprint arXiv:2011.00564*, 2020.

Maclay H., Osgood C. E., "Hesitation phenomena in spontaneous English speech", *Word*, vol. 15, nº 1, p. 19-44, 1959.

Mahl G. F., "Disturbances and silences in the patient's speech in psychotherapy.", *The Journal of Abnormal and Social Psychology*, vol. 53, nº 1, p. 1, 1956.

Mairesse F., Walker M. A., Mehl M. R., Moore R. K., "Using linguistic cues for the automatic recognition of personality in conversation and text", *Journal of artificial intelligence research*, vol. 30, p. 457-500, 2007.

Mehta Y., Majumder N., Gelbukh A., Cambria E., "Recent trends in deep learning based personality detection", *Artificial Intelligence Review*, vol. 53, nº 4, p. 2313-2339, 2020.

Meteer M. W., Taylor A. A., MacIntyre R., Iyer R., *Dysfluency annotation stylebook for the switchboard corpus*, University of Pennsylvania Philadelphia, PA, 1995.

Mielke S. J., Szlam A., Dinan E., Boureau Y.-L., "Reducing conversational agents' overconfidence through linguistic calibration", *Transactions of the Association for Computational Linguistics*, vol. 10, p. 857-872, 2022.

Mikolov T., Chen K., Corrado G., Dean J., "Efficient Estimation of Word Representations in Vector Space", *arXiv preprint arXiv:1301.3781*, 2013.

Moore J. D., Tian L., Lai C., "Word-level emotion recognition using high-level features", *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, p. 17-31, 2014.

Nicholson H. B. M., "Disfluency in dialogue: attention, structure and function", 2007.

O'Shaughnessy D., "Timing patterns in fluent and disfluent spontaneous speech", *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, IEEE, p. 600-603, 1995.

Oviatt S., "Predicting spoken disfluencies during human-computer interaction", *Computer Speech and Language*, vol. 9, nº 1, p. 19-36, 1995.

Park S., Shim H. S., Chatterjee M., Sagae K., Morency L.-P., "Computational Analysis of Persuasiveness in Social Multimedia: A Novel Dataset and Multimodal Prediction Approach",

Proceedings of the 16th International Conference on Multimodal Interaction, ICMI 2014, Association for Computing Machinery, New York, NY, USA, 2014.

Pennebaker J. W., Francis M. E., Booth R. J., "Linguistic inquiry and word count: LIWC 2001", *Mahway: Lawrence Erlbaum Associates*, vol. 71, nº 2001, p. 2001, 2001.

Pfeifer L. M., Bickmore T., "Should agents speak like, um, humans? The use of conversational fillers by virtual agents", *International Workshop on Intelligent Virtual Agents*, Springer, p. 460-466, 2009.

Purver M., Hough J., Howes C., "Computational Models of Miscommunication Phenomena", *Topics in Cognitive Science*, vol. 10, nº 2, p. 425-451, 2018.

Rasipuram S., Rao S. P., Jayagopi D. B., "Automatic prediction of fluency in interface-based interviews", *2016 IEEE Annual India Conference (INDICON)*, p. 1-6, 2016.

Rohanian M., Hough J., "Re-framing Incremental Deep Language Models for Dialogue Processing with Multi-task Learning", *CoRR*, 2020.

Rohanian M., Hough J., "Best of Both Worlds: Making High Accuracy Non-incremental Transformer-based Disfluency Detection Incremental", *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 3693-3703, 2021.

Rohanian M., Hough J., Purver M., "Alzheimer's Dementia Recognition Using Acoustic, Lexical, Disfluency and Speech Pause Features Robust to Noisy Inputs", *arXiv preprint arXiv:2106.15684*, 2021.

Ruder S., "ICLR 2021 Outstanding Papers, Char Wars, Speech-first NLP, Virtual conference ideas", Apr, 2020.

Salesky E., Sperber M., Waibel A., "Fluent Translations from Disfluent Speech in End-to-End Speech Translation", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, p. 2786-2792, June, 2019.

Schiffrin D., *Discourse markers*, nº 5, Cambridge University Press, 1987.

Schuller B., Weninger F., Zhang Y., Ringeval F., Batliner A., Steidl S., Eyben F., Marchi E., Vinciarelli A., Scherer K. *et al.*, "Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge", *Computer Speech & Language*, vol. 53, p. 156-180, 2019.

Sen P., "Speech Disfluencies occur at Higher Perplexities", *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, Association for Computational Linguistics, Online, p. 92-97, December, 2020.

Shalyminov I., Eshghi A., Lemon O., "Multi-Task Learning for Domain-General Spoken Disfluency Detection in Dialogue Systems", *CoRR*, 2018.

Shriberg E., "To 'errrr'is human: ecology and acoustics of speech disfluencies", *Journal of the International Phonetic Association*, vol. 31, nº 1, p. 153-169, 2001.

Shriberg E., Bates R. A., Stolcke A., "A prosody only decision-tree model for disfluency detection.", *Eurospeech*, vol. 97, Citeseer, p. 23832386, 1997.

Shriberg E. E., Preliminaries to a theory of speech disfluencies, PhD thesis, Citeseer, 1994.

Shriberg E. E., Phonetic consequences of speech disfluency, Technical report, SRI INTERNATIONAL MENLO PARK CA, 1999.

Shriberg E. E., Lickley R. J., "Intonation of clause-internal filled pauses", *Phonetica*, vol. 50, n° 3, p. 172-179, 1993.

Shriberg E., Stolcke A., Jurafsky D., Coccaro N., Meteer M., Bates R., Taylor P., Ries K., Martin R., Van Ess-Dykema C., "Can prosody aid the automatic classification of dialog acts in conversational speech?", *Language and speech*, vol. 41, n° 3-4, p. 443-492, 1998.

Skantze G., Johansson M., Beskow J., "Exploring turn-taking cues in multi-party human-robot discussions about objects", *Proceedings of the 2015 ACM on international conference on multimodal interaction*, p. 67-74, 2015.

Skidmore L., Moore R., "Incremental Disfluency Detection for Spoken Learner English", *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, Association for Computational Linguistics, Seattle, Washington, p. 272-278, July, 2022.

Stolcke A., Shriberg E., "Statistical Language Modeling for Speech Disfluencies", *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings (ICASSP)*, vol. 1, IEEE, p. 405-408, 1996.

Swerts M., "Filled Pauses as Markers of Discourse Structure", *Journal of Pragmatics*, vol. 30, n° 4, p. 485 - 496, 1998.

Székely É., Henter G. E., Beskow J., Gustafson J., "How to train your fillers: uh and um in spontaneous speech synthesis", *The 10th ISCA Speech Synthesis Workshop*, 2019.

Tian L., Moore J. D., Lai C., "Emotion recognition in spontaneous and acted dialogues", *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, p. 698-704, 2015.

Tottie G., "On the use of uh and um in American English", *Functions of Language*, vol. 21, n° 1, p. 6-29, 2014.

Tran T., Toshniwal S., Bansal M., Gimpel K., Livescu K., Ostendorf M., "Joint modeling of text and acoustic-prosodic cues for neural parsing", *arXiv preprint arXiv:1704.07287*, 2017.

Tran T., Yuan J., Liu Y., Ostendorf M., "On the Role of Style in Parsing Speech with Neural Models", *Proceedings of Interspeech 2019*, p. 4190-4194, 2019.

Tur G., De Mori R., *Spoken language understanding: Systems for extracting semantic information from speech*, John Wiley & Sons, 2011.

Vasilescu I., Rosset S., Adda-Decker M., "On the Role of Discourse Markers in Interactive Spoken Question Answering Systems", *LREC*, 2010.

Vinciarelli A., Mohammadi G., "A survey of personality computing", *IEEE Transactions on Affective Computing*, vol. 5, n° 3, p. 273-291, 2014.

Wang F., Chen W., Yang Z., Dong Q., Xu S., Xu B., "Semi-Supervised Disfluency Detection", *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, p. 3529-3538, August, 2018.

Wollermann C., Lasarcyk E., Schade U., Schröder B., "Disfluencies and uncertainty perception–evidence from a human–machine scenario", *Sixth Workshop on Disfluency in Spontaneous Speech*, 2013.

Yoshida E., Lickley R. J., "Disfluency patterns in dialogue processing", *DiSS-LPSS Joint Workshop 2010*, 2010.

Zayats V., Ostendorf M., "Giving attention to the unexpected: using prosody innovations in disfluency detection", *arXiv preprint arXiv:1904.04388*, 2019.

Zayats V., Tran T., Wright R., Mansfield C., Ostendorf M., "Disfluencies and Human Speech Transcription Errors", *arXiv preprint arXiv:1904.04398*, 2019.

Zhu G., Caceres J.-P., Salamon J., "Filler Word Detection and Classification: A Dataset and Benchmark", *arXiv preprint arXiv:2203.15135*, 2022.

Les corpus arborés avant et après le numérique

Sylvain Kahane* — **Nicolas Mazziotta****

* *Modyco, Université Paris Nanterre & CNRS*
** *U.R. Traverses, Université de Liège*

RÉSUMÉ. *Nous montrons comment, du* XVIII*e siècle à nos jours, l'annotation syntaxique de corpus a évolué de l'analyse exhaustive de phrases isolées à celle de listes d'exemples, puis à celle de textes entiers. Nous étudions l'évolution des visées de ces corpus arborés entre motivations pédagogique, théorique et ressources pour le TAL. Nous présentons quelques ouvrages clés, souvent peu connus de la communauté TAL comme de celle des linguistes : Buffier (1709), Beauzée (1765), Gaultier (1817), Clark (1847), Jespersen (1937) et Tesnière (1959). Nous concluons sur les liens actuels entre corpus arborés et TAL.*

MOTS-CLÉS : *corpus arboré, treebank, annotation syntaxique, analyse syntaxique, diagramme syntaxique.*

TITLE. *Treebanks before and after the digital technology*

ABSTRACT. *This paper explains how, from the 18th century to the present day, the syntactic annotation has evolved from the comprehensive analysis of isolated sentences to lists of examples, then to complete texts. We study the evolution of the aims of these treebanks between pedagogical and theoretical motivations and resources for NLP. We introduce some key works, often little known by the NLP community as well as by linguists: Buffier (1709), Beauzée (1765), Gaultier (1817), Clark (1847), Jespersen (1937), Tesnière (1959). We conclude on the current links between treebanks and NLP.*

KEYWORDS: *syntactic treebank, syntactic annotation, parsing, syntactic diagram.*

1. Introduction

Dans cet article[1], nous nous intéressons à l'histoire des corpus arborés en syntaxe. Par *corpus arborés*, nous entendons des corpus qui comprennent un certain nombre de phrases extraites de productions attestées auxquelles sont associées des analyses syntaxiques complètes. Ces analyses possèdent généralement une structure proche de celle d'un arbre de dépendance ou de constituants, d'où l'appellation commune de *corpus arboré* (angl. *treebank*). Si les corpus arborés se sont largement développés à l'âge numérique sous l'impulsion du traitement automatique des langues, nous allons montrer que ces ressources ont d'abord été développées à des fins pédagogiques, puis à des fins théoriques afin de valider les premiers modèles syntaxiques.

On situe généralement l'apparition des premiers corpus arborés dans les années 1970 avec le *Talbanken* du suédois (Einarsson, 1976), puis leur diffusion dans les années 1990 avec le *Penn Tree Bank* de l'anglais (Marcus *et al.*, 1993) et le *Prague Dependency Treebank* du tchèque (Hajič, 1998). Toutefois, il s'agit là des premières ressources au format numérique. Des ressources traditionnelles que nous considérons comme d'authentiques corpus arborés non numériques se sont en effet développées suite à l'apparition de démarches d'analyse syntaxique systématique d'exemples attestés au XVIII^e^ siècle avec Buffier (1709), puis de manière encore plus formalisée chez les encyclopédistes, Dumarsais (1754) et Beauzée (1765), cf. Kahane (2020). De nombreux ouvrages didactiques de grammaire du XIX^e^ siècle, à commencer par ceux du méconnu Louis Gaultier (1817), proposent de véritables collections d'exemples (et d'exercices corrigés) analysés systématiquement dans un même formalisme. Les analyses de Gaultier prennent la forme de diagrammes tabulaires qui représentent la structure de la phrase, dont l'un d'entre eux n'est pas sans rappeler le standard CoNLL (Buchholz et Marsi, 2006) (voir section 2). Au cours du XIX^e^ siècle se développent différentes conventions graphiques pour représenter la structure syntaxique et certains auteurs proposent des ouvrages entiers d'exemples analysés selon leurs conventions. Nous discuterons en particulier des ouvrages de Clark (1863) et de Reed et Kellogg (1889). C'est seulement au XX^e^ siècle que des linguistes davantage intéressés par les questions théoriques que didactiques s'emparent de la question et proposent des corpus d'exemples analysés dans le cadre théorique qu'ils défendent. C'est le cas en particulier de Jespersen (1937), de Tesnière (1959) et de Nida (1966).

Pour cadrer la discussion, nous commençons par un état de l'art succinct des corpus arborés à l'âge du numérique (section 2). Nous procédons ensuite chronologiquement. Nous commençons par l'étude des premières analyses syntaxiques systématiques proposées au XVIII^e^ siècle (section 3), car c'est cet intérêt pour l'exhaustivité qui a rendu possible la constitution de collections d'exemples analysés. Le cœur de l'article est consacré aux ouvrages pédagogiques du XIX^e^ siècle comportant d'importants corpus arborés dont les analyses sont présentées sous forme de diagrammes tabulaires ou hiérarchiques (section 4). Nous contrastons ces travaux avec ceux du XX^e^ siècle, qui en raison de leur visée théorique proposent généralement des corpus

1. Les deux auteurs ont contribué de manière équivalente à cette recherche.

multilingues (section 5). Notre conclusion se penche sur les pratiques actuelles et l'avenir des corpus arborés (section 6).

2. Les corpus arborés à l'âge du numérique

Une courte présentation des corpus arborés actuels permettra de mieux situer les travaux faits dans les siècles précédents. Si le premier corpus arboré électronique, le *Talbanken* du suédois (Einarsson, 1976), se développe dans les années 1970, c'est dans les années 1990 avec le *Penn Tree Bank* de l'anglais (Marcus *et al.*, 1993), que la communauté des linguistes et des talistes commence à s'intéresser vraiment à ce type de ressources. Le *Penn Tree Bank* a été développé sous l'impulsion de talistes avec l'utilisation d'outils de TAL pour la pré-annotation et dans l'objectif de développer des outils de TAL plus perfomants[2]. L'annotation syntaxique du corpus est une analyse en constituants encodée sous la forme d'un parenthésage du texte. Conformément au principe générativiste de mouvement, l'arbre syntaxique inclut des nœuds vides, comme le nœud *-NONE-*, dans la figure 1, qui indique la position du groupe syntaxique extrait *how many credit cards*, auquel il est lié par un index (cf. l'index *1* dans *WHNP-1* et **T*-1*)[3].

```
( (SBARQ
    (INTJ (UH So) )
    (WHNP-1
      (WHADJP (WRB how) (JJ many) )
      (, ,)
      (INTJ (UH um) )
      (, ,) (NN credit) (NNS cards) )
    (SQ (VBP do)
      (NP-SBJ (PRP you) )
      (VP (VB have)
        (NP (-NONE- *T*-1) )))
    (. ?) (-DFL- E_S) ))
```

Figure 1. *Exemple extrait du Penn Tree Bank :* So how many, um, credit cards do you have ?

2. On peut citer les premières phrases de l'introduction de Marcus *et al.* (1993) : « *There is a growing consensus that significant, rapid progress can be made in both text understanding and spoken language understanding by investigating those phenomena that occur most centrally in naturally occurring unconstrained materials and by attempting to automatically extract information about language from very large corpora. Such corpora are beginning to serve as an important research tool for investigators in natural language processing, speech recognition, and integrated spoken language systems, as well as in theoretical linguistics.* »

3. Exemple extrait de la page `catalog.ldc.upenn.edu/desc/addenda/LDC99T42.mrg.txt` du catalogue LDC qui distribue le Penn Tree Bank.

Le développement du *Prague Dependency Treebank* (Hajič, 1998) a débuté à la suite du *Penn Tree Bank*. Il s'agit d'un corpus de tchèque annoté en syntaxe de dépendance, avec une couche d'annotation en syntaxe de surface (appelée *analytical tree*) accompagnée d'une analyse en syntaxe profonde (appelée *tectogrammatical tree*), ainsi qu'un outil d'édition et de visualisation des arbres, comme le montrent les arbres de la figure 2 (Hajic *et al.*, 2001).

Plusieurs corpus arborés sont développés à la suite de ces premières expériences, avec une grande variété de schémas d'annotation et de formats d'encodage, jusqu'à la proposition du format tabulaire CoNLL[4]. Ce format particulièrement économique, inspiré du format proposé un an plus tôt par Hall et Nivre (2006), est aujourd'hui un standard pour l'encodage des analyses en dépendance sur des corpus de textes (Buchholz et Marsi, 2006). Il a contribué à populariser l'analyse en dépendance dans le domaine du TAL et tout particulièrement de l'analyse syntaxique automatique.

La figure 3 illustre l'encodage au format CoNLL d'un extrait du corpus arboré SUD_French-GSD[5]. Dans cet encodage, la structure est décrite dans un tableau généralement à 10 colonnes. Les mots de la phrase sont dans la colonne 2. La colonne 1 contient leur identifiant, la colonne 3 les lemmes, la colonne 4 les parties du discours, la colonne 6 les traits morphosyntaxiques standard et la colonne 10 des traits additionnels. L'arbre de dépendance est encodé dans les colonnes 7 et 8 : la colonne 7 contient l'identifiant du gouverneur de chaque mot (avec un 0 pour le mot 3 qui n'a pas de gouverneur) et la colonne 8 sa fonction syntaxique. Par exemple, le mot 1 est le *det* du mot 2. Les colonnes 5 et 9 restent vides (elles sont utilisées par les parsers) et la colonne 10 est un fourre-tout d'informations additionnelles.

De nombreux outils permettent le requêtage et l'affichage de fichiers au format CoNLL, comme Grew-match (Guillaume, 2021), sur lequel nous reviendrons dans la section 6, ou SETS (Luotolahti *et al.*, 2015). Certains outils d'annotation permettent de modifier dynamiquement un CoNLL à partir de sa forme graphique, comme ArboratorGrew (Gerdes, 2013 ; Guibon *et al.*, 2020), UD Annotatrix (Tyers *et al.*, 2017) ou ConnluEditor (Heinecke, 2019).

Maintenant que ce cadre est posé, nous pouvons entamer notre retour vers le futur en commençant par les analyses syntaxiques du début du XVIII[e] siècle.

4. D'après le colloque éponyme en apprentissage automatique, *Conference in Natural Language Learning*.

5. Le format SUD (Surface-Syntactic Universal Dependencies, `surfacesyntacticud.github.io`, (Gerdes *et al.*, 2018)) est une variante du format UD (Universal Dependencies, `universaldependencies.org`, (Nivre *et al.*, 2016 ; **?**)), où, contrairement à UD, les mots fonctionnels sont traités comme des têtes.

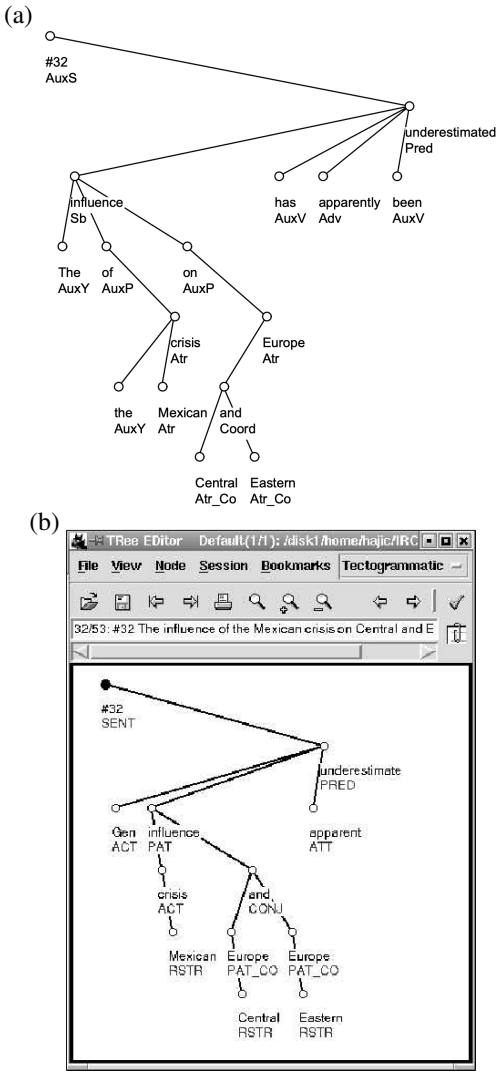


Figure 2. *Arbres analytique (a) et tectogrammatique (b) :* The influence of the Mexican crisis on Central and Eastern Europe has apparently been underestimated

3. Premières analyses syntaxiques exhaustives au XVIII^e siècle

Notre parcours historique sur les corpus arborés syntaxiques commence au XVIII^e siècle, où l'on trouve les premières analyses syntaxiques complètes de phrases

```
1   Le        le        DET     _   Definite=Def[...]   2   det        _   [...]
2   roi       roi       NOUN    _   Gender=Masc[...]    3   subj       _   _
3   pouvait   pouvoir   VERB    _   Mood=Ind[...]       0   root       _   _
4   le        le        PRON    _   Gender=Masc[...]    5   comp:obj   _   _
5   croire    croire    VERB    _   VerbForm=Inf        3   comp:obj   _   [...]
6   .         .         PUNCT   _   _                   3   punct      _   _
```

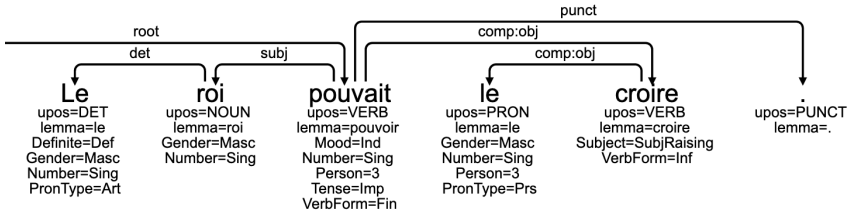


Figure 3. *Exemple extrait du SUD_French-GSD:* Le roi pouvait le croire *(fr-ud-train_01347; certaines informations omises sont indiquées par « [...] »)*

attestées[6]. Nous reproduisons ici l'intégralité d'une analyse du jésuite Claude Buffier (1661-1737), tirée de sa grammaire du français (1709, 84)[7] :

> « *Un homme qui étourdit les gens qu'il rencontre avec de frivoles discours a coutume de causer beaucoup d'ennui à tout le monde*. Je dis que dans ce discours, tous les mots sont pour modifier le nom *un homme*, et le verbe *a coutume*, et que c'est en cela que consiste tout le mystère et toute l'essence de la syntaxe des langues : 1° le nom *un homme*, est modifié d'abord par le *qui* déterminatif : car il ne s'agit pas ici d'un homme en général, mais d'un homme marqué et déterminé en particulier par l'action qu'il fait d'*étourdir* ; de même, il ne s'agit pas d'un homme *qui étourdit* en général, mais *qui étourdit* en particulier *les gens*, et non pas *les gens* en général, mais en particulier *les gens qu'il rencontre*. Or cet homme qui étourdit ceux qu'il rencontre, est encore particularisé par *avec des discours*, et *discours* est encore particularisé par *frivoles*. On peut voir le même dans la suite de la phrase : *a coutume* est particularisé par *de causer*, *de causer* est particularisé par ses deux régimes, par son régime absolu, savoir, *beaucoup d'ennui*, et par son régime respectif, *à tout le monde*. Voilà donc comment tous les mots d'une phrase quelque longue qu'elle soit, ne sont que pour modifier le nom et le verbe. »

Bien que la distinction entre la syntaxe et la sémantique ne soit pas encore aboutie, les termes *modifier*, *déterminer* et *particulariser* peuvent être compris comme « dé-

6. On trouve bien sûr des analyses syntaxiques avant cela, mais, à notre connaissance, jamais aussi complètes et systématiques. On pourra notamment consulter la remarquable grammaire de l'anglais (1653) que John Wallis (1616-1703) a rédigée en latin – traduction anglaise par Kemp (1972). Voir Imrényi et Mazziotta (2020) pour un historique des analyses en dépendance depuis Priscien.

7. L'orthographe et les mises en italiques sont modernisées.

pendre de » ou « être complément de ». Le terme *régime* correspond au terme *complément*, qui ne sera véritablement introduit que par Beauzée (voir plus loin).

Nous proposons de représenter notre interprétation de l'analyse de Buffier par le diagramme de la figure 4, où les flèches expriment les relations du type « est déterminé par », « est particularisé par » ou « est modifié par ». Notons que les deux termes de la relation sont à chaque fois assez clairement donnés par Buffier, chaque élément mentionné dans le texte particularisant le précédent. On peut voir que même si la terminologie comme l'argumentation ne distinguent pas clairement syntaxe et sémantique, la description est totalement compatible avec une analyse syntaxique actuelle.

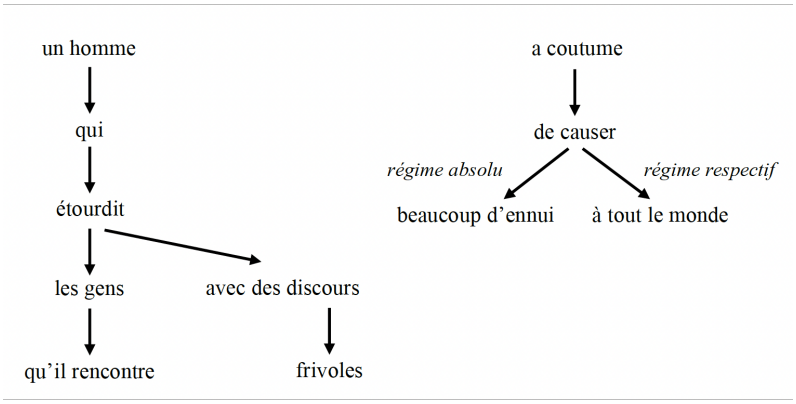


Figure 4. *Diagramme réalisé par nos soins à partir de l'analyse de Buffier (1709) :* Un homme qui étourdit les gens qu'il rencontre avec de frivoles discours a coutume de causer beaucoup d'ennui à tout le monde

On trouve d'autres analyses de phrases complexes chez d'autres auteurs français au XVIII^e^ siècle, notamment chez Girard (1747) et dans l'*Encyclopédie* – dans les articles de Dumarsais (1754) et de Beauzée (1765). Voir Kahane (2020) pour une étude critique.

On doit à Beauzée la notion moderne de *complément* (Chevalier, 1968). Dans l'article de l'*Encyclopédie* qu'il consacre au terme *Régime*, Beauzée introduit un sous-article *Complément* (Beauzée n'a été en charge des articles de linguistique de l'*Encyclopédie* qu'à partir de la lettre *F*). Plus précisément, Beauzée distingue le *complément grammatical* ou *initial*, qui est un mot, du *complément logique* ou *total*, qui en est la projection, combinant ainsi les notions modernes de dépendance et de constituance[8] :

8. L'orthographe de la citation qui suit est modernisée.

> « Par exemple, dans cette phrase, *avec les soins requis dans les circonstances de cette nature* ; le mot *nature* est le complément grammatical de la préposition *de* : *cette nature* en est le complément logique : la préposition *de* est le complément initial du nom appellatif *les circonstances* ; et *de cette nature* en est le complément total : *les circonstances*, voilà le complément grammatical de la préposition *dans* ; et *les circonstances de cette nature* en est le complément logique. [...] » (Beauzée, 1765, 5)

Comme nous l'avons fait pour l'analyse de Buffier, nous pouvons proposer une diagrammatisation de l'analyse de Beauzée ou plus exactement des deux analyses superposées proposées par Beauzée. Dans la figure 5, nous représentons les relations « être le complément initial ou grammatical » par des bulles et « être le complément total ou logique » par des flèches.

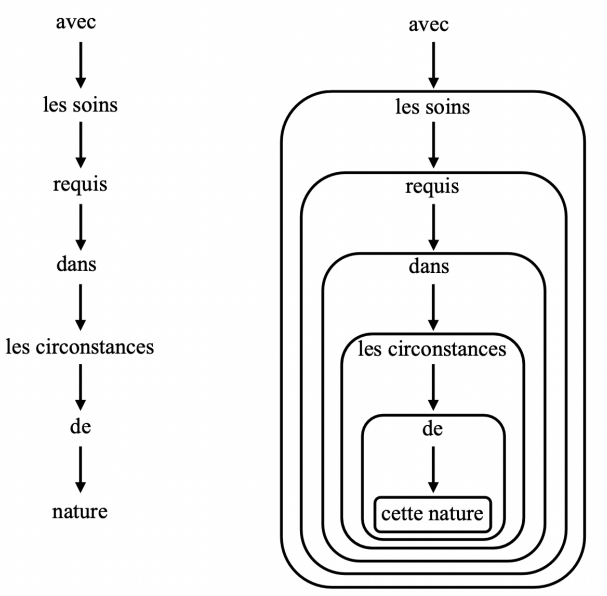


Figure 5. *Diagrammes réalisés par nos soins à partir de l'analyse double de Beauzée :* avec les soins requis dans les circonstances de cette nature

La démarche d'analyse syntaxique sous la forme d'un discours où tous les mots sont décrits successivement se développe particulièrement dans la grammaticographie anglo-saxonne, sous forme de parsing. À titre d'exemple, on peut citer cet extrait des exercices de Lindlay Murray (1745-1826), qui fait autorité dans le domaine de l'enseignement de la grammaire à la fin du XVIII^e^ siècle et au début du suivant :

> « "He who lives virtuously prepares for all events." *He* is a personal pronoun, of the third person singular number and masculine gender. *Who* is a

> relative pronoun, which has for its antecedent “he,” with which it agrees in gender and number [...]. *Lives* [is] a regular verb neuter, indicative mood, present tense, third person singular, agreeing with its nominative, “who” [...]. *Virtuously* is an adverb of quality. *Prepares* [is] a regular verb neuter, indicative mood, present tense, third person singular, agreeing with its nominative, “he.” *For* is a preposition. *All* is an adjective pronoun, of the indefinite kind, the plural number, and belongs to its substantive, “events,” with which it agrees [...]. *Events* is a common substantive of the neuter gender, the third person, in the plural number, and the objective case, governed by the preposition, “for” [...]. » (Murray, 1812, 18)

On voit que ces discours permettent de décrire une partie de la structuration syntaxique de phrases complètes, mais ils portent sur des phrases isolées et sont présentés d'une manière qui ne favorise pas les comparaisons. On peut encore ajouter, comme nous l'a fait remarquer un des relecteurs, que les exemples peuvent être éparpillés dans le texte et ne constituent pas, de ce fait, un corpus strict, c'est-à-dire un matériau textuel présenté indépendamment de son exégèse. Le recours ultérieur aux diagrammes va radicalement changer les choses.

4. Des corpus arborés pour la didactique au XIX^e^ siècle

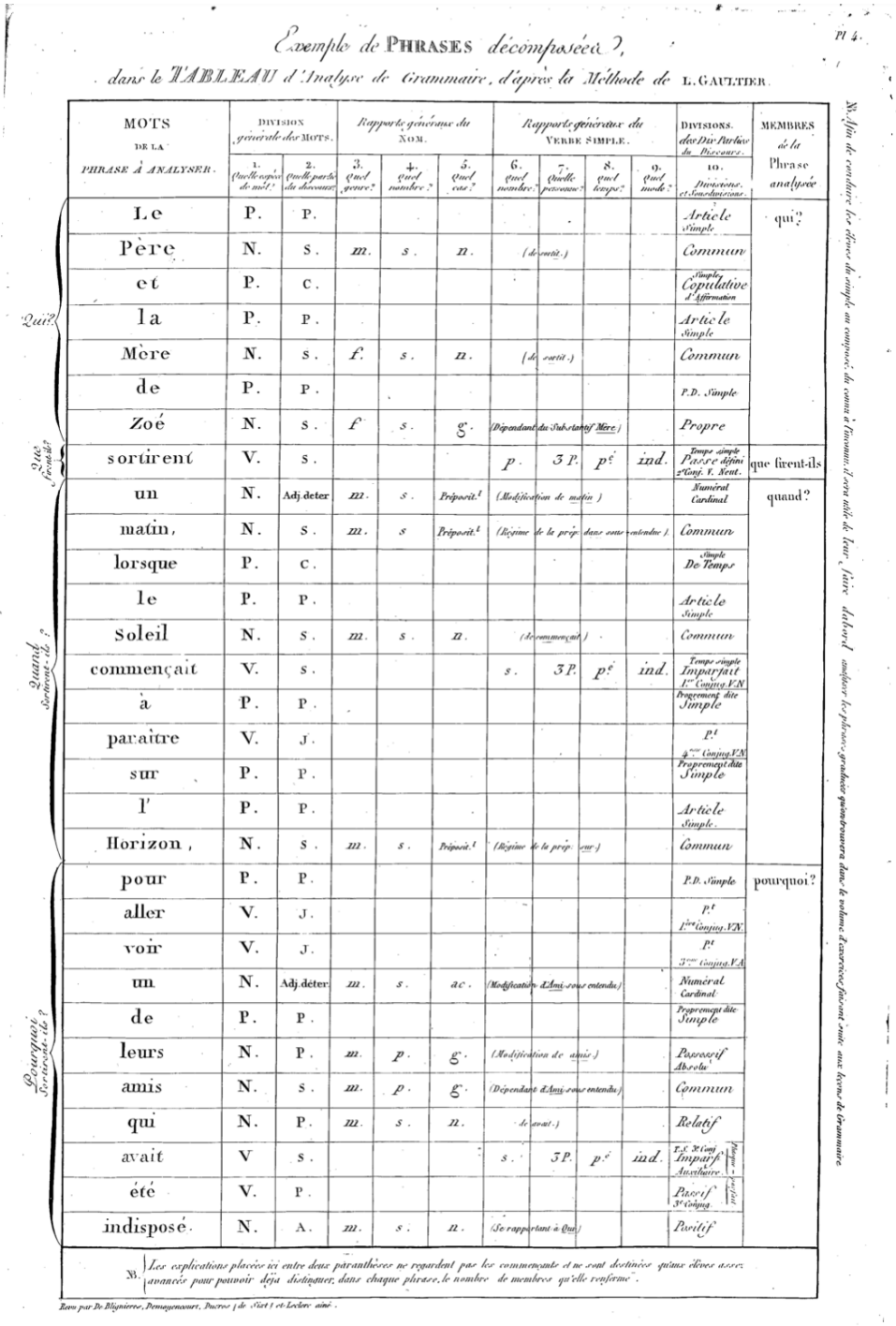
L'idée que l'on peut analyser des structures de manière exhaustive se développe au XIX^e^ siècle de deux façons : premièrement, on voit apparaître de grandes listes d'exemples analysés ; deuxièmement et probablement conséquemment, les analyses prennent la forme de diagrammes, les rendant à la fois plus concises et plus lisibles. Ces listes d'analyses, qui constituent indéniablement de véritables corpus arborés, sont développées au sein d'ouvrages didactiques monolingues visant à présenter les différentes constructions d'une langue donnée. Nous aborderons successivement le cas des diagrammes tabulaires, en particulier ceux de Louis Gaultier (sous-section 4.1) et des diagrammes hiérarchiques proposés par Clark et Reed et Kellogg (sous-section 4.2).

4.1. *Premiers diagrammes tabulaires*

À notre connaissance, le premier auteur à utiliser extensivement des diagrammes pour l'enseignement de la grammaire est l'abbé Louis Gaultier (1746-1818), dont l'*Atlas de grammaire* (1817) contient une grande variété de diagrammes à visée pédagogique[9]. Il comporte en particulier un diagramme tabulaire, qui encode un fragment d'analyse en dépendance (figure 6).

Dans l'introduction des *Élémens de grammaire* publié en 1829 par ses élèves, on trouve un diagramme similaire précédé de la description suivante :

9. Les numérotations de pages qui suivent renvoient à la copie pdf distribuée par `gallica.bnf.fr`, qui comprend différents feuillets numérotés séparément.



Exemple de PHRASES décomposées,

dans le TABLEAU d'Analyse de Grammaire, d'après la Méthode de L. GAULTIER.

Pl 4.

MOTS DE LA PHRASE À ANALYSER.	DIVISION générale des MOTS. 1. Quelle espèce de mot?	2. Quelle partie du discours	Rapports généraux du NOM. 3. Quel genre?	4. Quel nombre?	5. Quel cas?	Rapports généraux du VERBE SIMPLE. 6. Quel nombre?	7. Quelle personne?	8. Quel temps?	9. Quel mode?	DIVISIONS. des Dix Parties du Discours. 10. Divisions et Subdivisions.	MEMBRES de la Phrase analysée
Le	P.	P.								Article Simple	qui?
Père	N.	S.	m.	s.	n.	(devenu.)				Commun	
et	P.	C.								Simple Copulative d'Affirmation	
la	P.	P.								Article Simple	
Mère	N.	S.	f.	s.	n.	(devenu.)				Commun	
de	P.	P.								P.D. Simple	
Zoé	N.	S.	f	s.	g.	(Dépendant du Substantif Mère.)				Propre	
sortirent	V.	S.				p.	3 P.	p.é	ind.	Temps simple Passé défini 2 Conj. V. Neut.	que firent-ils
un	N.	Adj. déter	m.	s.	Prépositif	(Modification de matin.)				Numéral Cardinal	quand?
matin,	N.	S.	m.	s	Prépositif	(Régime de la prép. dans sous entendue.)				Commun	
lorsque	P.	C.								Simple De Temps	
le	P.	P.								Article Simple	
Soleil	N.	S.	m.	s.	n.	(Le commençait.)				Commun	
commençait	V.	S.				s.	3 P.	p.é	ind.	Temps simple Imparfait 1 Conj. V.N	
à	P.	P.								Progressif dite Simple	
paraître	V.	J.								P.t 4.ème Conj. V.N.	
sur	P.	P.								Proprement dite Simple	
l'	P.	P.								Article Simple	
Horizon,	N.	S.	m.	s.	Prépositif	(Régime de la prép. sur.)				Commun	
pour	P.	P.								P.D. Simple	pourquoi?
aller	V.	J.								P.t 1.ère Conj. V.N.	
voir	V.	J.								P.t 3.ème Conj. V.A.	
un	N.	Adj. déter	m.	s.	ac.	(Modificatif d'Ami sous entendu.)				Numéral Cardinal	
de	P.	P.								Proprement dite Simple	
leurs	N.	P.	m.	p.	g.	(Modification de amis.)				Possessif Absolu	
amis	N.	S.	m.	p.	g.	(Dépendant d'Ami sous entendu.)				Commun	
qui	N.	P.	m.	s.	n.	(Le avait.)				Relatif	
avait	V	S.				s.	3 P.	p.é	ind.	2.e 3.e Conj. Imparf. Auxiliaire.	Temps composé
été	V.	P.								Passif 3.e Conjug.	
indisposé.	N.	A.	m.	s.	n.	(Se rapportant à Qui.)				Passif	

NB. Les explications placées ici entre deux parenthèses ne regardent pas les commençants et ne sont destinées qu'aux élèves assez avancés pour pouvoir déjà distinguer, dans chaque phrase, le nombre de membres qu'elle renferme.

Qui? Fait quoi? Quand? Sortirent-ils? Pourquoi? Sortirent-ils?

NB. Afin de rendre les élèves du simple au composé, du connu à l'inconnu, il sera utile de leur faire d'abord simplifier les phrases, qu'ensuite ils retrouveront dans le volume d'exercices, qui sont suite aux leçons de Grammaire

Rem par De Blignieres, Pennequin court, Pacros fils de Siret et Leclerc aîné.

Figure 6. *Reproduction d'un tableau d'analyse grammaticale par Gaultier (1817, 11) :* Le Père et la Mère de Zoé sortirent un matin, lorsque le Soleil commençait à paraître sur l'Horizon, pour aller voir un de leur amis qui avait été indisposé.

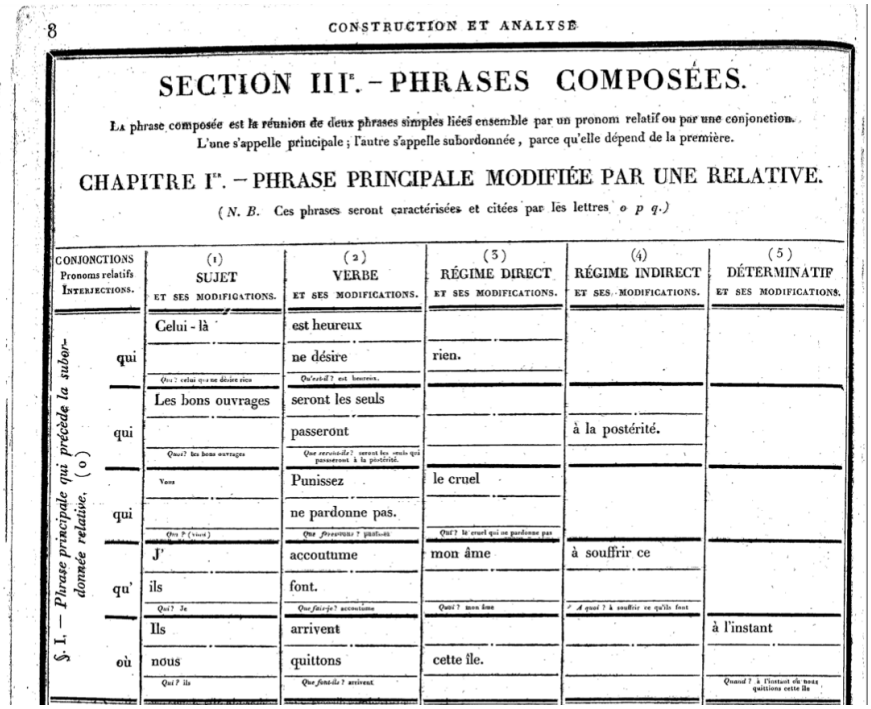
> « Pour faire l'analyse grammaticale, il faut avoir une feuille de papier, une ardoise ou un tableau noir partagé en dix colonnes. Dans une marge à gauche, on écrira les mots de la phrase à analyser les uns au-dessous des autres. Dans la première colonne, on indiquera à laquelle des trois parties primitives du discours, et dans la seconde à laquelle des dix parties secondaires du discours chaque mot appartient ; dans la troisième, la quatrième et la cinquième, on marquera le genre, le nombre et le cas des noms ; dans la sixième, la septième, la huitième et la neuvième, on indiquera le nombre, la personne, le temps en général et le mode du verbe personnel. Dans la dixième, on indiquera toutes les divisions et les subdivisions des dix parties du discours. »

Le diagramme de Gaultier de la figure 6 s'apparente fortement au format CoNLL utilisé aujourd'hui. On trouve en particulier, dans la colonne 5 intitulée *Quel cas ?*[10], un encodage des dépendances pour les noms de la phrase : ainsi *Père* et *Mère* sont analysés comme sujet (*n.* [pour nominatif] de *sortirent*), *Zoé* comme un dépendant génitival (*g.* de *Mère*), *matin* comme un complément (*régime de la préposition* dans *sous-entendue*), etc. La terminologie de Gaultier est plus traditionnelle que celles de ses prédécesseurs encyclopédistes (section 3), dont Beauzée, qui avait distingué précisément la notion morphosyntaxique de régime de la notion syntaxique de complément. Ainsi, les relations syntaxiques sont à nouveau encodées par des noms de cas (« nominatif », « génitif », etc.).

Même si on peut supposer qu'il a été utilisé à plusieurs reprises avec des élèves, ce premier type de diagramme reste sporadique dans les ouvrages de Gaultier et de ses étudiants. En revanche, un autre type de diagramme tabulaire totalise près de 200 exemples analysés dans Gaultier (1817, 17-36), et quelques-uns dans de Blignières *et al.* (1829, 228-244). Nous reproduisons dans la figure 7 quelques exemples de phrases comportant des propositions relatives. Dans cette analyse, six positions syntaxiques sont considérées : complémenteur, sujet, verbe, complément d'objet direct (régime direct), complément oblique (régime indirect) et complément circonstanciel (déterminatif)[11]. Chaque proposition est divisée en segments positionnés les uns à la suite des autres dans ces six positions. Par exemple, si l'on prend le dernier exemple de la figure 7, « Ils arrivent à l'instant où nous quittons cette île. » (Gaultier, 1817, 34),

10. La colonne 5 ne contient que le cas lui-même, mais celui-ci est complété par un texte entre parenthèses qui déborde sur les colonnes suivantes normalement consacrées aux catégories de la forme verbale. On notera, tout en bas du tableau, la mention suivante : *Les explications placées ici entre deux parenthèses ne regardent pas les commençants et ne sont destinées qu'aux élèves assez avancés pour pouvoir déjà distinguer dans chaque phrase le nombre de membres qu'elle renferme.* Les « membres » de la phrase en question sont indiqués dans la dernière colonne (non numérotée), où la phrase est découpée en quatre segments identifiés par autant de questions : *qui ? que firent-ils ? quand ? pourquoi ?* Cette segmentation et ces questions figurent aussi sur le côté gauche du tableau sous forme d'accolades.

11. On trouve déjà chez Girard (1747) des analyses syntaxiques de ce type – voir l'étude de Kahane (2020, 113-120, en particulier 118) –, mais elles ne sont pas utilisées de manière aussi systématique que chez Gaultier (1817).



8 CONSTRUCTION ET ANALYSE

SECTION III^e. – PHRASES COMPOSÉES.

La phrase composée est la réunion de deux phrases simples liées ensemble par un pronom relatif ou par une conjonction. L'une s'appelle principale ; l'autre s'appelle subordonnée, parce qu'elle dépend de la première.

CHAPITRE I^er. – PHRASE PRINCIPALE MODIFIÉE PAR UNE RELATIVE.

(*N. B.* Ces phrases seront caractérisées et citées par les lettres *o p q.*)

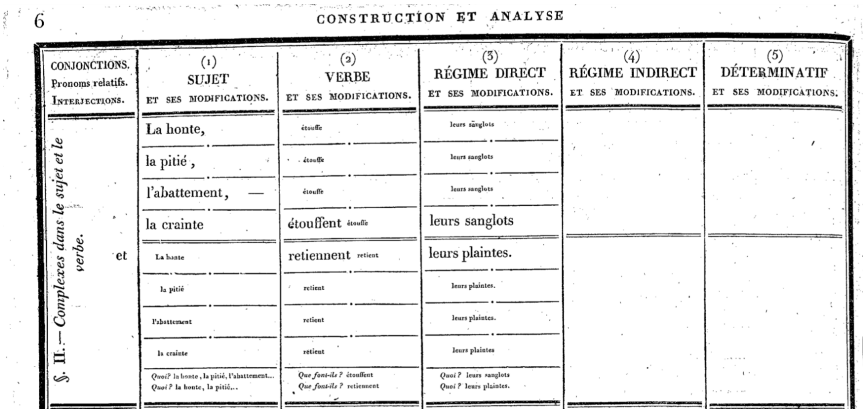
CONJONCTIONS Pronoms relatifs Interjections.	(1) SUJET ET SES MODIFICATIONS.	(2) VERBE ET SES MODIFICATIONS.	(3) RÉGIME DIRECT ET SES MODIFICATIONS.	(4) RÉGIME INDIRECT ET SES MODIFICATIONS.	(5) DÉTERMINATIF ET SES MODIFICATIONS.
§. I. — *Phrase principale qui précède la subordonnée relative.* (o)	Celui-là	est heureux			
qui		ne désire	rien.		
	Qui? celui qui ne désire rien	Qu'est-il? est heureux.			
	Les bons ouvrages	seront les seuls			
qui		passeront		à la postérité.	
	Quoi? les bons ouvrages	Que deviendra-t-il? seront les seuls qui parviendront à la postérité.			
	Vous	Punissez	le cruel		
qui		ne pardonne pas.			
	Qui? (vous)	Que fera-vous? pardonner	Qui? le cruel qui ne pardonne pas		
	J'	accoutume	mon âme	à souffrir ce	
qu'	ils	font.			
	Qui? Je	Que fait-je? accoutume	Quoi? mon âme	A quoi? à souffrir ce qu'ils font	
	Ils	arrivent			à l'instant
où	nous	quittons	cette île.		
	Qui? ils	Que fait-elle? arrivent			Quand? à l'instant où nous quittons cette île

Figure 7. *Analyse de phrases complexes chez Gaultier (1817, 34)*

l'analyse indique que *ils* | *arrivent* | *à l'instant* se décompose en sujet-verbe-modifieur et la relative *où* | *nous* | *quittons* | *cette île* en complémenteur-sujet-verbe-objet. Le fait que la relative forme un constituant avec *à l'instant* est indiqué dans la troisième ligne de l'analyse (« *Quand ?* à l'instant où nous quittons cette île »).

Les analyses tabulaires que l'on rencontre pour la première fois chez Gaultier se développent dans différentes langues – en particulier en langue allemande par Becker (1829), puis sous une autre forme dans les éditions ultérieures selon Hudson, puis, sans doute sous l'influence de cette dernière, en langue anglaise, notamment dans les grammaires de Morell (1852) et de Meiklejohn (1886)[12]. La volonté des auteurs est toujours de « faire voir » la logique de l'analyse. Les colonnes formalisent le typage d'éléments récurrents. Elles permettent donc incidemment de retrouver des occurrences (tokens) de catégories grammaticales (types). L'analyse de la coordination qui en découle est particulièrement intéressante (voir figure 8) : elle rappelle les analyses en grille proposées par Blanche-Benveniste *et al.* (1979) avec la disposition

12. Nous remercions Richard Hudson pour les discussions au sujet de Gaultier, Becker, Morell et Meiklejohn. Des matériaux issus des grammaires de ces auteurs sont présentés sur son site (`dickhudson.com/uk/`).



6 CONSTRUCTION ET ANALYSE

CONJONCTIONS. Pronoms relatifs. INTERJECTIONS.	(1) SUJET ET SES MODIFICATIONS.	(2) VERBE ET SES MODIFICATIONS.	(3) RÉGIME DIRECT ET SES MODIFICATIONS.	(4) RÉGIME INDIRECT ET SES MODIFICATIONS.	(5) DÉTERMINATIF ET SES MODIFICATIONS.
§. II. — Complexes dans le sujet et le verbe.	La honte,	étouffe	leurs sanglots		
	la pitié,	étouffe	leurs sanglots		
	l'abattement, —	étouffe	leurs sanglots		
	la crainte	étouffent étouffe	leurs sanglots		
et	La honte	retiennent retient	leurs plaintes.		
	la pitié	retient	leurs plaintes.		
	l'abattement	retient	leurs plaintes.		
	la crainte	retient	leurs plaintes		
	Quoi? la honte, la pitié, l'abattement... *Quoi ?* la honte, la pitié...	*Que font-ils ?* étouffent *Que font-ils ?* retiennent	*Quoi ?* leurs sanglots *Quoi ?* leurs plaintes.		

Figure 8. *Analyse de coordinations par Gaultier (1817, 20)*

verticale des paradigmes entre conjoints, à la différence que Gaultier complète l'analyse pour mettre en évidence que les paradigmes de 4 et 2 éléments se combinent pour donner $4 \times 2 = 8$ propositions élémentaires[13].

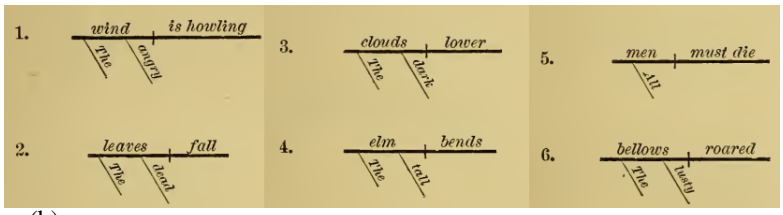
4.2. *Les premiers diagrammes hiérarchiques et les* Keys

Les premiers diagrammes hiérarchiques apparaissent dans les années 1830 dans une grammaire du latin (1832) par le grammairien allemand Johann Gustav Freidrich Billroth (1808-1836) (unique diagramme connu d'un auteur mort prématurément) et dans une grammaire de l'anglais à destination des sourds (1836) par le savant américain Frederick A. P. Barnard (1809-1889)[14]. Les données n'y sont pas représentées sous la forme de tableaux, mais sous celle d'un réseau d'éléments hiérarchisés. À partir de 1847, Stephen W. Clark propose une série de grammaires de l'anglais comportant un grand nombre d'exemples analysés par des diagrammes arborescents originaux. La naissance de ces diagrammes syntaxiques hiérarchiques ouvre en effet la possibilité de collectionner des listes d'analyses exhaustives. Les grammaires de Stephen W. Clark (1810-1901) (Clark, 1847 ; Clark, 1855) et d'Alonzo Reed (?-1899) et Brainerd Kellogg (1834-1920) (Reed et Kellogg, 1876 ; Reed et Kellogg, 1877) sont ainsi accompagnées d'ouvrages qualifiés de « *Keys* », c'est-à-dire de solutions aux exercices (Clark, 1863 ; Reed et Kellogg, 1889). Dans ces ouvrages, les phrases

13. Comme remarqué par Kahane (2012), Tesnière (1959) propose, comme dans l'analyse en grille, de traiter les relations entre conjoints orthogonalement aux relations de subordination et il met en évidence la combinaison des paradigmes dans ses stemmas 265 et 266, p. 345.

14. Voir la thèse peu diffusée de Brittain (1973), ainsi que l'étude de Mazziotta et Kahane (2017), qui décrit les propriétés des premiers diagrammes d'analyse en constituants.

(a)



(b)

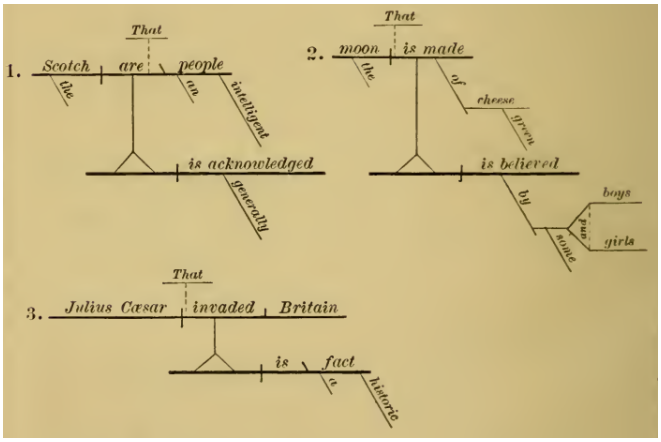


Figure 9. *Liste d'analyses diagrammatiques : (a) Reed et Kellogg, 1889 : 1 ; (b) Reed et Kellogg, 1889 : 30*

d'exercices de la grammaire correspondante (qui peuvent être inventées ou tirées de la littérature) sont accompagnées d'une analyse exhaustive sous la forme d'un diagramme[15,16]. La figure 9 donne deux exemples tirés de la *Key* de Reed et Kellogg (1889), qui s'appuie sur les grammaires des auteurs (Reed et Kellogg, 1876 ; Reed et Kellogg, 1877).

15. Nous renvoyons à Mazziotta (2016) concernant les systèmes de Clark et à Gleason (1965, 142-161) pour une présentation succincte du système de Reed et Kellogg, encore en usage de nos jours – voir, par exemple, le manuel de Otto et Bauer (2019). Chez Reed et Kellogg, chaque mot correspond à un trait, horizontal ou oblique selon qu'il s'agit d'un nom ou verbe ou d'un modifieur. Clark associe, quant à lui, chaque mot à une bulle. La structure de la phrase est la combinaison directe de ces traits ou de ces bulles, sans que les relations entre les mots ne soient représentées par un signe discret.

16. La démarche est ici radicalement opposée aux habitudes traditionnelles comme celle de Murray (1799), qui est plutôt une correction d'exercices d'identification d'erreurs orthographiques ou grammaticales. La *Key* de Murray correspondant à la 16e édition (Murray, 1812) ne comporte pas de correction des exercices d'analyse morphosyntaxique que l'édition précédente du livre d'exercices comporte (section 3).

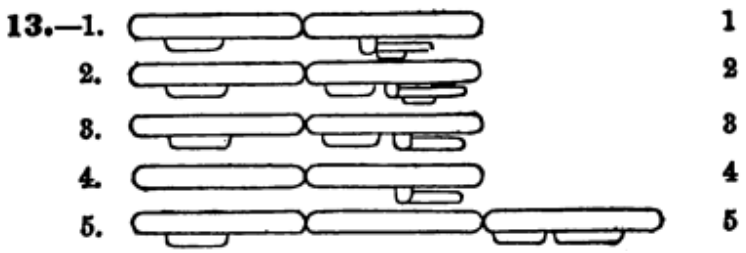


Figure 10. *Liste d'analyses diagrammatiques sans étiquetage des entités réifiant les mots (Clark 1863 : 10) : 1.* The sun rose on the sea*, 2.* A mist rose slowly from the lake*, 3.* The night passed away in song*, 4.* Morning retrurned in joy *et 5.* The mountains showed their gray heads *(Clark 1863 : 11, cf. Clark 1855 : 13)*

Les diagrammes sont regroupés en fonction du point de grammaire que chaque exemple illustre. La mise en série a comme effet de faire ressortir graphiquement la partie commune partagée par tous les diagrammes. Dans la figure 9a, on observe un trait horizontal et un petit trait vertical qui l'interrompt, ainsi que des traits obliques sous la première partie du trait horizontal ; dans la figure 9b, on repère le dessin en forme de « Y » inversé. La mise en série de diagrammes met en évidence un invariant graphique qui correspond à l'illustration d'un point de grammaire identifié dans les ouvrages auxquels se réfère la *Key*. Dans la figure 9a, il s'agit de la construction des « modified subjects » (Reed et Kellogg, 1876, (1880) 27-30). Dans la figure 9b, il s'agit des « noun clauses » (Reed et Kellogg, 1876, (1880) 81-82). L'exemple de Clark (1863), qui s'appuie sur Clark (1855), est plus précoce, mais plus complexe. Il n'étiquette pas les mots dans ses listes de diagrammes. Ce sont des phrases réelles qui figurent en vis-à-vis de la figure 10, qui représente l'analyse.

De la même manière qu'expliqué pour la *Key* de Reed et Kellogg, les diagrammes ont tous comme point commun de comporter un invariant (ici l'agencement horizontal d'au moins deux bulles). Tant Clark que Reed et Kellogg ont conscience que l'analyse syntaxique complète force à prendre position (Clark, 1863, 3) et pousse à trancher dans certains cas naturellement ambigus (Reed et Kellogg, 1889, s.n.). Il s'agit là d'une conséquence directe de la recherche d'exhaustivité, dont on sait les implications pour la constitution des corpus à l'heure actuelle. La perspective des auteurs est en quelque sorte renversée par rapport à la nôtre : eux voient effectivement leurs *Keys* comme des solutions aux exercices proposés dans d'autres ouvrages (dont ils sont la « clé »), alors que nous considérons ces *Keys* comme des corpus annotés dont les grammaires constituent le point d'entrée. Il est évident que le contraste principal avec les corpus modernes est que ces derniers permettent aux utilisateurs de définir eux-mêmes des requêtes, alors que les anciens ouvrages (format papier oblige) constituent une sorte d'index d'un nombre limité de requêtes. Il en résulte que la sélection des requêtes représentées correspond à ce que les pédagogues ont jugé pertinent de mettre à disposition de leurs lecteurs, selon une progression pédagogique qui correspond à

celle de leur grammaire et non dans la visée théorique des auteurs qui suivront (section 5) ou dans la visée exploratoire des outils d'exploitation des corpus actuels (6). Une autre différence se situe au niveau de la saillance de ce qui est représenté. Les *Keys* ne mettent en évidence les structures que de manière indirecte (il faut chercher l'invariant graphique entre les exemples). De leur côté, les outils modernes permettent la mise en évidence d'un pivot focal dans les résultats de la requête (par exemple en surlignant les mots).

La démarche d'accumulation d'exemples à des fins pédagogiques ouvre la voie à l'exploitation théorique de listes similaires.

5. Valider une théorie par des corpus arborés au XX^e^ siècle

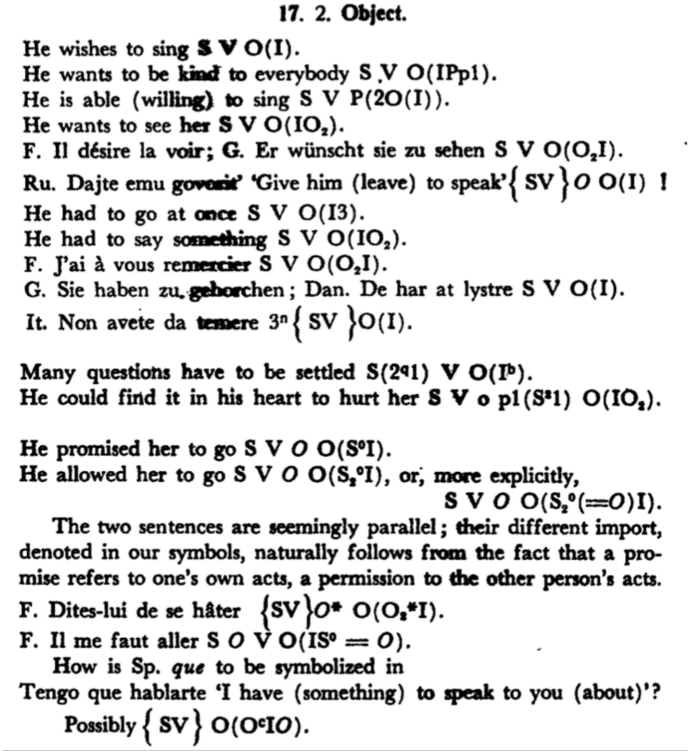
Les ouvrages du XIX^e^ siècle comportant des corpus d'exemples arborés sont des grammaires à visée pédagogique. Elles portent sur une langue unique, le français pour Gaultier, l'anglais pour Clark ou Reed et Kellogg. Les structures sont utilisées pour présenter les différentes constructions de la langue qu'il s'agit d'apprendre, sans qu'un cadre théorique général, applicable à différentes langues, ne soit dégagé. C'est au XX^e^ siècle que se développent les premiers ouvrages de syntaxe générale, dont une particularité importante est qu'ils comportent des exemples de plusieurs langues[17]. Nous observons les démarches d'Otto Jespersen (sous-section 5.1) et de Lucien Tesnière (5.2) pour illustrer notre propos.

5.1. *Otto Jespersen*

Après avoir publié son grand ouvrage de linguistique générale, *Philosophy of grammar* (Jespersen, 1924), Otto Jespersen (1860-1943) propose l'ouvrage intitulé *Analytic syntax* (Jespersen, 1937), qui est une collection organisée d'exemples dans plusieurs langues européennes (anglais, latin, danois, français, espagnol, portugais, italien, finnois, allemand, russe et grec). Il y développe un système de notation original permettant de diagrammatiser par une formule la structure syntaxique de chaque exemple étudié (sur ce système, voir Cigana, 2020, 232-234).

Il s'agit essentiellement d'une analyse en constituants, comme le souligne James D. McCawley dans sa préface de l'édition de 1984 publiée par l'University of Chicago Press. Par exemple, dans la figure 11, la phrase *He wants to see her* est analysée S V O(IO_2), indiquant que dans cette structure de type SVO, O se décompose lui-même en un infinitif I et un objet O_2. Dans d'autres analyses de ce même extrait,

17. Nous devons mentionner ici la thèse de Weil (1844), exceptionnelle à de nombreux égards. Dans ce travail consacré à l'ordre des mots, Henri Weil (1818-1909) s'intéresse à plusieurs langues (latin, grec ancien, français, anglais, allemand, turc et chinois) et montre que la linéarisation est guidée par trois types de facteurs : la structure syntaxique (il considère une structure de dépendance à la suite de Beauzée et distingue les langues et constructions à têtes finales *vs* initiales), la prosodie et la structure thème-rhème.



17. 2. Object.

He wishes to sing S V O(I).
He wants to be kind to everybody S V O(IPp1).
He is able (willing) to sing S V P(2O(I)).
He wants to see her S V O(IO$_2$).
F. Il désire la voir; G. Er wünscht sie zu sehen S V O(O$_2$I).
Ru. Dajte emu govorit' 'Give him (leave) to speak' { SV } O O(I) !
He had to go at once S V O(I3).
He had to say something S V O(IO$_2$).
F. J'ai à vous remercier S V O(O$_2$I).
G. Sie haben zu gehorchen; Dan. De har at lystre S V O(I).
It. Non avete da temere 3^n { SV } O(I).

Many questions have to be settled S(2^{q}1) V O(I^b).
He could find it in his heart to hurt her S V o p1(S^{s}1) O(IO$_2$).

He promised her to go S V O O(S^{0}I).
He allowed her to go S V O O(S$_2$^{0}I), or, more explicitly,
S V O O(S$_2$^{0}(=O)I).
The two sentences are seemingly parallel; their different import, denoted in our symbols, naturally follows from the fact that a promise refers to one's own acts, a permission to the other person's acts.
F. Dites-lui de se hâter {SV} O* O(O$_2$*I).
F. Il me faut aller S O V O(IS^0 = O).
How is Sp. *que* to be symbolized in
Tengo que hablarte 'I have (something) to speak to you (about)'?
Possibly { SV } O(O^{c}IO).

Figure 11. *Section sur les infinitifs en position objet de Jespersen (1937, 48-49)*

Jespersen introduit des éléments vides, comme S^0 (pour *He promised her to go*), ou la coindexation d'un trace, notée *, avec une autre position (pour *Dites-lui de se hâter*). On pourra consulter, en contraste, l'interprétation dépendancielle de Cigana (2020) des annotations de Jespersen.

Dans son introduction, Jespersen souligne l'importance de l'exemple par la maxime suivante (non attribuée) : *Nur das Beispiel führt zum Licht; Vieles Reden tut es nicht* (Seul l'exemple permet de faire la lumière ; Les longs discours ne suffisent pas). Il poursuit en précisant la nature des exemples étudiés :

> « Certains sont empruntés à mes lectures, d'autres aux grammaires que l'on peut trouver couramment ; mais un grand nombre sont en fait des phrases préfabriquées. [...] je n'ai peut-être pas réussi à éviter une certaine monotonie [...] Ma seule excuse est que cet ouvrage n'a aucune prétention pédagogique [...] il importait surtout de trouver des phrases courtes et non ambiguës qui permettent d'illustrer l'emploi et l'utilité des formules dans un grand nombre de constructions grammaticales. » (Jespersen, 1937, (1971) 16-17)

Le travail à visée non pédagogique de Jespersen reste isolé dans cette première moitié du XXe siècle. Il faut attendre Nida (1966) ou Ross (1967) pour voir à nouveau des ouvrages comprenant de longues listes d'exemples syntaxiquement analysés dans une perspective théorisante.

5.2. *Lucien Tesnière*

Les *Éléments de syntaxe structurale*, ouvrage posthume de Lucien Tesnière (1893-1954) commencé en 1932 et publié en 1959, sont connus pour introduire un modèle théorique complet d'analyse en dépendance. Le livre contient également des matériaux procédant de la même démarche. Plusieurs analyses exhaustives de textes sous forme de diagrammes (dits « stemmas ») figurent à la fin de l'ouvrage (1959, 638-653)[18] : deux poèmes – *La cigale et la fourmi* de La Fontaine (voir la figure 12)[19] et *Le vase brisé* de Sully Prudhomme –, une longue phrase en grec de Platon et une autre de Tacite en latin, des extraits du *Polyeucte* et du *Cid* de Corneille, d'*Athalie* de Racine, de *Booz endormi* de Victor Hugo et du *Crime de Sylvestre Bonnard* d'Anatole France. Si nous citons la liste de ces textes, c'est qu'il s'agit, à notre connaissance, du premier exemple de corpus arborés basé sur des textes suivis attestés. Tous les travaux mentionnés jusque-là contenaient uniquement des analyses de phrases isolées, souvent construites ou simplifiées. Ces exemples sont précédés d'un chapitre intitulé « Le stemma intégral » (Tesnière, 1959, 629-32), dont voici quelques extraits :

> « 1. – Si nous faisons usage de toutes les possibilités que la stemmatisation d'une phrase peut nous offrir pour en représenter graphiquement l'infinie complication structurale, nous aboutissons à un stemma d'une complexité telle que nous n'y avons pratiquement à peu près jamais recouru au cours de cet ouvrage.
>
> 2. – Mais à côté des stemmas partiels et fragmentaires que nous avons utilisés pour faire comprendre telle ou telle partie de la syntaxe structurale, il est possible, au moins théoriquement, de concevoir un stemma intégral faisant état de tous les éléments structuraux rencontrés dans une phrase, ou tout au moins de se rapprocher de cet idéal. [...]
>
> 5 – Pratiquement nous n'avons guère eu l'occasion de présenter de stemmas de cette nature, le souci de la clarté de notre exposé nous ayant au

18. La tradition philologique de ces diagrammes n'est pas claire : dans l'introduction des *Éléments*, Fourquet indique qu'ils « ont été redessinés par M. Georges Bichet » (Tesnière, 1959, iv).

19. Indiquons quelques conventions utilisées par Tesnière dans ses stemmas. Les dépendances sont indiquées par des traits pleins, obliques pour les relations tête-dépendant et horizontaux pour la coordination. Les traits hachurés indiquent des relations de coréférence. Les symboles en forme de « T » indiquent un cas particulier de combinaison que Tesnière nomme la translation : ainsi dans le deuxième stemma, *de* translate *mouche* pour lui permettre d'occuper une position normalement dévolue à un adjectif. Les ronds pointés indiquent un translatif zéro. On pourra consulter Kahane et Osborne (2015) pour une analyse critique de l'ouvrage de Tesnière.

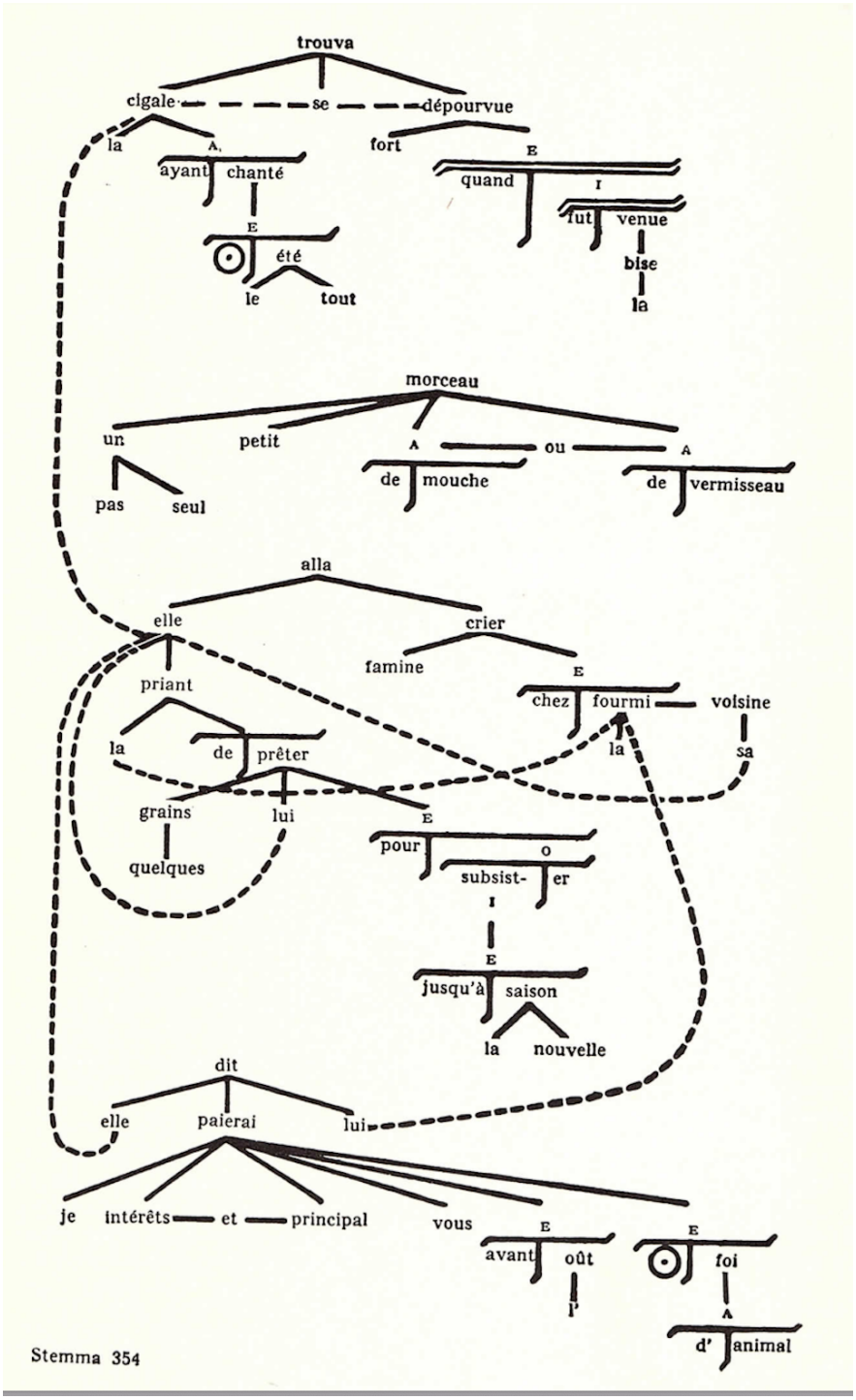


Figure 12. *Première moitié de l'analyse de* La cigale et la fourmi *par Tesnière (1959 : 638)*

> contraire toujours fait une loi de ne pas compliquer le fait à faire comprendre de faits superfétatoires. [...]
>
> 10. – Mais en réalité le langage ne se présente pas à nous comme une succession de phrases isolées. La normale est au contraire le plus souvent une suite de phrases qui expriment en principe des idées agencées entre elles de façon à former un tout organisé en vue d'exprimer, soit oralement, soit par écrit, une pensée plus ou moins complexe.
>
> 11. – [...] Le stemma intégral d'une interminable conversation ou d'un long discours comporte celui de toutes les phrases qui les composent. »

Tesnière souligne à la fois l'utilité d'adopter des exemples simples pour présenter le cadre théorique et la nécessité, au moment voulu, de confronter la théorie avec de vrais textes dans toute leur complexité.

Nous allons maintenant revenir sur la période actuelle à la lumière de notre parcours des deux siècles précédents.

6. Conclusion : des corpus arborés au XXIe siècle pour quel objectif

Les analyses de textes telles que celles proposées par Tesnière à la fin de son ouvrage sont difficiles à exploiter pour extraire de l'information. Elles servent à illustrer la théorie et, bien que l'on ne doutera pas de la visée pédagogique de l'auteur, elles nous paraissent surtout utiles à l'auteur, qui vérifie que son système lui permet d'appréhender n'importe quel énoncé. Des décennies après cet auteur, les systèmes de requêtes et de concordanciers, que le passage au numérique a permis, rendent possible de sélectionner toutes les constructions d'un type donné. Il devient alors également possible de combiner les avantages d'une analyse suivie d'un texte complet avec un classement pédagogique des phénomènes, comme dans les grammaires du XIXe siècle (section 4). Le corpus arboré peut ainsi être appréhendé selon différentes facettes, chaque requête offrant une vue particulière sur les données.

On a vu que les corpus analysés en syntaxe ont une longue histoire avant le numérique. Ils apparaissent sous une double pression : une pression pédagogique, pour donner aux apprenants d'une langue des exemples des différentes constructions de la langue (sections 3 et 4) ; une pression théorique (section 5), pour vérifier que les modèles de langue proposés ont une couverture exhaustive des phénomènes rencontrés. Avec l'apparition du numérique, les corpus sont devenus électroniques et l'annotation syntaxique a subi de nouvelles contraintes : paradoxalement, la dématérialisation de l'encodage a entraîné une simplification des annotations. Il est plus simple sur un texte informatique d'ajouter des parenthèses que de tracer des traits et des flèches (voir l'exemple du *Penn Tree Bank* de la figure 1). Il faudra attendre les années 2000 pour que se développe un moyen simple d'encoder des arbres de dépendance, le format tabulaire CoNLL (voir l'exemple de la figure 4). Mais un tel format est peu iconique, peu ergonomique pour des humains qui doivent parcourir la table en passant d'un identifiant à l'autre et n'est véritablement lisible qu'avec une interface proposant une repré-

sentation graphique sous forme d'un diagramme arborescent. Chacun de ces formats d'encodage, parenthésage ou format tabulaire, contraint fortement les utilisateurs : même si toutes sortes d'informations peuvent être encodées quelque part, certaines le sont plus facilement que d'autres, ce qui a conduit à un certain appauvrissement des représentations utilisées[20]. D'un autre côté, la numérisation a permis à l'annotation de connaître un nouvel essor : les données servent à présent à la fois d'input et d'output aux outils de TAL.

Aujourd'hui, le lien entre traitement automatique des langues et corpus arborés est complexe. Si au moment de l'avènement des corpus arborés numériques, l'apprentissage d'analyseurs syntaxiques automatiques sur des corpus arborés[21] apparaissait comme l'outil le plus puissant pour obtenir des modèles de langue, le développement actuel des méthodes d'analyse distributionnelle sur de très grands corpus, comme les *transformers* BERT (Devlin *et al.*, 2019), rend l'analyse syntaxique souvent dispensable. Les corpus arborés se trouvent alors renvoyés à leur usage initial, qui est celui d'une source d'exemples pour la pédagogie et pour l'étude théorique et la recherche de constructions nouvelles. Notons tout de même que, même s'ils ne jouent plus nécessairement un rôle central dans le développement à proprement parler des outils de traitement automatique des langues, les corpus arborés restent utiles pour l'évaluation des outils et pour vérifier que ceux-ci ont bien saisi la structure des énoncés.

À l'heure actuelle, ce sont donc davantage les développeurs et utilisateurs de corpus arborés qui ont besoin du traitement automatique des langues que l'inverse. Les méthodes d'apprentissage permettent un développement accéléré des corpus arborés : après avoir annoté manuellement quelques phrases, il est déjà possible d'apprendre un analyseur donnant des résultats suffisamment bons pour pré-annoter automatiquement le reste du corpus. En répétant régulièrement cette opération (procédure dite de *bootstrapping* ; Breiman, 1996 ; Seraji *et al.*, 2012), le processus d'annotation devient à chaque itération plus performant. On peut donner pour exemple les résultats d'une expérience faite par Guiller (2020) sur le corpus arboré SUD-Naija_NSC avec un parser bi-affine utilisant un BERT mutilingue[22]. Comme le montre la figure 13, avec 10 phrases, on dépasse 80 % de précision pour la reconnaissance des parties du discours (POS) et avec 100 phrases on a plus de 85 % d'étiquettes fonctionnelles correctes (LUS) et 75 % de gouverneurs reconnus (UAS)[23]. Ces résultats sont rendus possibles par l'utilisation d'un transformer BERT et donc d'un apprentissage préalable sur un

20. Parallèlement, les théoriciens du langage, dans un souci de formalisation, se sont eux-mêmes astreints à utiliser des structures mathématiques bien définies, ce qui les a amenés aussi à privilégier des structures d'arbres au détriment de structures plus ambitieuses. Voir par exemple, les arbres de constituants à la base de tous les modèles génératifs depuis Chomsky (1957).

21. La littérature sur l'apprentissage automatique d'analyseurs à partir de corpus arborés est immense. On pourra consulter Kübler *et al.* (2009) pour les principes de base.

22. Le naija est un pidgin-créole de l'anglais et le corpus Naija_NSC utilise l'orthographe standard de l'anglais pour les mots lexicaux. Les résultats seraient certainement moins bons avec une langue sans lien avec les langues ayant servi à l'entraînement du BERT multilingue.

23. Le LAS (*Labelled Attachment Score*) calcule le nombre de mots qui ont à la fois le bon gouverneur (UAS) et la bonne étiquette fonctionnelle (LUS).

modèle	POS	LUS	UAS	LAS
1 phrases	51.32	41.18	16.14	10.07
10 phrases	82.13	72.95	39.53	33.72
100 phrases	93.38	86.17	75.28	68.15
1000 phrases	97.29	93.44	90.92	86.46
5000 phrases	97.89	94.73	93.39	89.48

Figure 13. *Performances atteintes par un parser bi-affine basé sur un BERT multilingue et entraîné sur des tailles différentes du SUD-Naija_NSC (Guiller 2020, 51)*

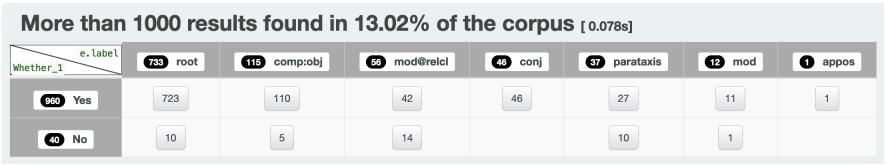


Figure 14. *Requête Grew-match avec double clustering*

grand corpus brut – voir également les expériences de de Lhoneux *et al.* (2022). Il est aussi possible d'utiliser des méthodes d'apprentissage par transfert si l'on possède des corpus arborés d'autres langues ayant des constructions similaires (Aufrant, 2018).

Concernant l'utilisation des corpus arborés, les systèmes de requêtes permettent d'accéder aux données et de faire différents tris. La plateforme Grew-match (Guillaume, 2021) permet notamment de regrouper les résultats d'une requête selon différentes clés. La figure 14 donne le résultat d'une requête[24] qui nous dit si oui ou non le sujet `S` d'un verbe `V` se trouve avant `V` (`whether_1`) en fonction de la position syntaxique de `V` (`e.label`). On voit par exemple que quand le verbe `V` est la tête d'une relative (`mod@relcl`), il y a 14 sujets inversés contre 42 dans l'ordre standard. En cliquant sur le 14, on obtient les exemples correspondants.

Associés à de tels outils, les corpus arborés deviennent des instruments puissants pour l'analyse et la description d'une langue. Le développement linéaire de la base Universal Dependencies depuis 2014 (environ 25 treebanks et 15 langues supplémen-

24. La requête est `pattern { e: G−>V; V −[subj]−> S; S[upos=NOUN] }`. Autrement dit, on recherche un motif (`pattern`) du graphe avec deux nœuds `V` et `S`, où `S` est un sujet nominal de `V`. Le lien `e` est introduit pour pouvoir interroger la fonction de `V`, c'est-à-dire la relation qui le lie à son gouverneur `G`. Les résultats sont triés d'abord en fonction de la position `S` par rapport à `V` (`S << V`; réponse « Yes » ou « No ») et de l'étiquette de la relation entre `V` et son gouverneur `G` (`e.label`). (`universal.grew.fr/?custom=62cdcb8a55a6b`).

taires chaque année), ainsi que la diversité toujours plus grande des langues qui bénéficient d'un corpus arboré, montre un intérêt soutenu de l'ensemble de la communauté des linguistes, des linguistes de terrain aux talistes, en passant par les théoriciens. Le développement de corpus arborés, initié par des linguistes pour la compréhension de la grammaire, l'enseignement et la validation des théories linguistiques, puis boosté par la communauté TAL de l'avènement du numérique à aujourd'hui, se trouve ainsi à nouveau investi par les linguistes avec des possibilités de développement et d'exploitation démultipliées par les travaux en TAL.

Remerciements

Nous souhaitons remercier Loïc Grobol pour ses commentaires sur la première version de ce travail, Richard Hudson pour les échanges à propos des grammaires anciennes, ainsi que les trois relecteurs de la revue pour leurs très nombreuses remarques.

7. Bibliographie

Aufrant L., Training parsers for low-resourced languages : improving cross-lingual transfer with monolingual knowledge, PhD thesis, Université Paris Saclay, 2018.

Barnard F. A. P., *Analytic grammar, with symbolic illustration*, French, New York, 1836.

Beauzée N., « Régime », *in* D. Diderot, J. L. R. D'Alembert (eds), *Encyclopédie*, vol. 14, p. 5-11, 1765.

Becker K. F., *Deutsche Grammatik*, J. C. Hermann'sche, Frankfurt, 1829.

Billroth J. G. F., *Lateinische Syntax für die obern Klassen gelehrter Schulen*, Weidmann, 1832.

Blanche-Benveniste C., Borel B., Deulofeu J., Durand J., Giacomi A., Loufrani C., « Des grilles pour le français parlé », *Recherches sur le Français Parlé*, n° 2, p. 163-206, 1979.

Breiman L., « Bagging predictors », *Machine learning*, vol. 24, n° 2, p. 123-140, 1996.

Brittain R. C., A critical history of systems of sentence diagramming in English, PhD thesis, University of Texas, Austin, 1973.

Buchholz S., Marsi E., « CoNLL-X shared task on multilingual dependency parsing », *Proceedings of the tenth Conference on Computational Natural Language Learning (CoNLL)*, p. 149-164, 2006.

Buffier C., *Grammaire françoise sur un plan nouveau*, Le Clerc-Brunet-Leconte & Montalant, Paris, 1709.

Chevalier J.-C., *Histoire de la syntaxe : Naissance de la notion de complément dans la grammaire française (1530–1750)*, Droz, Paris, 1968.

Chomsky N., *Syntactic structures*, Mouton, 1957.

Cigana L., « Some aspects of dependency in Otto Jespersen's structural syntax », *in* A. Imrényi, N. Mazziotta (eds), *Chapters of Dependency Grammar : A historical survey from antiquity to Tesnière*, John Benjamins, Amsterdam/Philadelphia, p. 215-251, 2020.

Clark S. W., *The science of the English grammar : A practical grammar in which words, phrases, and sentences are classified to their offices, and their relation to each other, illustrated by a complete system of diagrams*, H. W. Barnes & Company, Cincinnati, 1847.

Clark S. W., *The science of English language. A practical grammar [...]. Revised edition*, A.S. Barnes & Co., Derby, Bradley & Co., New York, 1855.

Clark S. W., *Key to Clark's grammar : in which the analyses of the sentences of the grammar are indicated by diagrams*, A.S. Barnes & Burr, New York, 1863.

de Blignières, Demoyencourt, Ducrot (de Sixt), Le Clerc ainé, *Élemens de grammaire française, extraits de la grammaire de l'abbé Gaultier*, Jules Renouard, Paris, 1829.

de Lhoneux M., Zhang S., Søgaard A., « Zero-Shot Dependency Parsing with Worst-Case Aware Automated Curriculum Learning », *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, Association for Computational Linguistics, Dublin, Ireland, p. 578-587, 2022.

Devlin J., Chang M.-W., Lee K., Toutanova K., « Bert : Pre-training of deep bidirectional transformers for language understanding », 2019.

Dumarsais C. C., « Construction », *in* D. Diderot, J. L. R. D'Alembert (eds), *Encyclopédie*, vol. 4, p. 73-92, 1754.

Einarsson J., « Talbankens skriftspråkskonkordans », 1976.

Gaultier L., *Atlas de grammaire, ou tables propres à exciter et à soutenir l'attention des enfans dans l'étude de cette science*, Jules Renouard, Paris, 1817.

Gerdes K., « Collaborative dependency annotation », *Proceedings of the second international conference on dependency linguistics (DepLing)*, p. 88-97, 2013.

Gerdes K., Guillaume B., Kahane S., Perrier G., « SUD or surface-syntactic universal dependencies : An annotation scheme near-isomorphic to UD », *Proceedings of the second Universal Dependencies Workshop (UDW)*, Association for Computational Linguistics (ACL), 2018.

Girard G., *Les vrais principes de la langue françoise ou la parole réduite en méthode*, Le Breton, Paris, 1747.

Gleason H. A. J., *Linguistics and English grammar*, Holt, Rinehart and Winston, New York, Chicago, San Francisco, Toronto and London, 1965.

Guibon G., Courtin M., Gerdes K., Guillaume B., « When collaborative treebank curation meets graph grammars », *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 2020.

Guillaume B., « Graph Matching and Graph Rewriting : GREW tools for corpus exploration, maintenance and conversion », *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : System Demonstrations*, p. 168-175, 2021.

Guiller K., *Analyse syntaxique du pidgin-créole du Nigéria à l'aide d'un transformer (BERT) : Méthodes et résultats*, Université Sorbonne Nouvelle, 2020. Mémoire de master.

Hajič J., « Building a syntactically annotated corpus : The prague dependency treebank », *in* E. Hajičová (ed.), *Issues of valency and meaning : Studies in honour of Jarmila Panevová*, Karolinum, p. 106-132, 1998.

Hajic J., Vidová-Hladká B., Pajas P., « The prague dependency treebank : Annotation structure and support », *Proceedings of the IRCS workshop on linguistic databases*, p. 105-114, 2001.

Hall J., Nivre J., « A generic architecture for data-driven dependency parsing », *Proceedings of the 15th Nordic Conference of Computational Linguistics (NODALIDA 2005)*, p. 47-56, 2006.

Heinecke J., « ConlluEditor : A fully graphical editor for Universal Dependencies treebank files », *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest)*, p. 87-93, 2019.

Imrényi A., Mazziotta N., *Chapters of Dependency Grammar : A historical survey from antiquity to Tesnière*, nº 212 in *Studies in Language Companion Series*, John Benjamins, Amsterdam/Philadelphia, 2020.

Jespersen O., *The philosophy of language*, Allen & Unwin, Londres, 1924.

Jespersen O., *Analytic syntax*, Allen & Unwin, Londres, 1937.

Kahane S., « De l'analyse en grille à la modélisation des entassements », *in* S. Caddeo, M.-N. Roubaud, M. Rouquier, F. Sabio (eds), *Penser les langues avec Claire Blanche-Benveniste*, Presses Universitaires de Provence, p. 101-116, 2012.

Kahane S., « How dependency syntax found its modern form in the French Encyclopedia : From Buffier (1709) to Beauzée (1765) », *in* A. Imrényi, N. Mazziotta (eds), *Chapters of Dependency Grammar : A historical survey from antiquity to Tesnière*, John Benjamins, Amsterdam/Philadelphia, p. 85-131, 2020.

Kahane S., Osborne T., « Translators' introduction », *Elements of structural syntax*, John Benjamins, Amsterdam/Philadelphia, p. xxix-lxxiv, 2015.

Kemp J. A., *John Wallis's grammar of the English language*, Longman, London, 1972.

Kübler S., McDonald R., Nivre J., « Dependency parsing », *Synthesis Lectures on Human Language Technologies*, vol. 1, nº 1, p. 1-127, 2009.

Luotolahti J., Kanerva J., Pyysalo S., Ginter F., « SETS : Scalable and efficient tree search in dependency graphs », *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) : Demonstrations*, p. 51-55, 2015.

Marcus M., Santorini B., Marcinkiewicz M. A., « Building a large annotated corpus of English : The Penn Treebank. 1993 », *Computational linguistics*, 1993.

Mazziotta N., « Drawing sentences before syntactic trees : Stephen Watkins Clark's sentence diagrams (1847) », *Historiographia linguistica*, vol. 43, nº 3, p. 301-342, 2016.

Mazziotta N., Kahane S., « To what extent is immediate constituency analysis dependency-based ? A survey of foundational texts », *Proceedings of the fourth international conference on Dependency Linguistics (Depling)*, ACL, p. 116-126, 2017.

Murray L., *A key to the exercises : adapted to L. Murray's English grammar*, Longman & Rees, Darton & Harvey et Wilson, Spence & Mawman, London, 1799.

Murray L., *English exercises, adapted to Murray's English grammar : [...]*, 16 edn, Collins & Co., New York, 1812.

Nida E., *A synopsys of English syntax*, Mouton and Co, London/The Hague, 1966.

Nivre J., De Marneffe M.-C., Ginter F., Goldberg Y., Hajic J., Manning C. D., McDonald R., Petrov S., Pyysalo S., Silveira N. *et al.*, « Universal dependencies v1 : A multilingual treebank collection », *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, p. 1659-1666, 2016.

Otto J., Bauer S. W., *The diagramming dictionary : A complete reference tool for young writers, aspiring rhetoricians, and anyone else who needs to understand how English works*, Well-Trained Mind Press, Charles City, Virginia, 2019. OCLC : 1104510563.

Reed A., Kellogg B., *Graded lessons in English. An elementary English grammar [...]*, Clark and Maynard, New York, 1876.

Reed A., Kellogg B., *Higher lessons in English. A work on grammar and composition [...]*, Clark and Maynard, New York, 1877.

Reed A., Kellogg B., *A key containing diagrams of the sentences given for analysis in Reed and Kellogg's Graded lessons in English and Higher lessons in English*, Effingham Maynard & Co., New York, 1889.

Seraji M., Megyesi B., Nivre J., « Bootstrapping a Persian dependency treebank », *Linguistic Issues in Language Technology*, 2012.

Tesnière L., *Éléments de syntaxe structurale*, Klincksieck, Paris, 1959.

Tyers F., Sheyanova M., Washington J., « UD Annotatrix : An annotation tool for Universal Dependencies », *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT)*, p. 10-17, 2017.

Wallis J., *Grammatica Linguae Anglicanae*, Leon Lichfield, Oxford, 1653. [Grammar of the English Language].

Weil H., De l'ordre des mots dans les langues anciennes comparées aux langues modernes, PhD thesis, Sorbonne, Paris, 1844.

Notes de lecture

Rubrique préparée par Denis Maurel

Université de Tours, LIFAT (Laboratoire d'informatique fondamentale et appliquée)

Sylvain KAHANE, Kim GERDES. Syntaxe théorique et formelle. Volume 1 : Modélisation, unités, structures. *Language science press*. 2022. 627 pages. ISBN : 978-3-98554-037-2.

Lu par **Quentin FELTGEN**

Ghent University, Belgique

Avec cet ouvrage ambitieux, les deux auteurs offrent un modèle qui permet de décrire les mécanismes sous-jacents à la production des énoncés linguistiques, depuis la structuration du sens jusqu'à l'ordonnancement linéaire des unités linguistiques, l'une et l'autre étant articulés par une double structure syntaxique, profonde et de surface. Le contenu de l'ouvrage ne se limite cependant pas à l'exposé d'un modèle linguistique donné ; les auteurs décrivent également le cadre épistémologique plus général dans lequel cette théorie s'inscrit et peut être scientifiquement évaluée. En ce sens, ils situent la syntaxe des dépendances comme un modèle possible répondant à des problèmes de recherche bien définis, modèle auquel il existe donc des alternatives. Cette démarche invite à la fois à la réflexion et à la discussion, et d'autres théories linguistiques (comme la syntaxe X-barre ou la grammaire des constructions) sont convoquées à l'occasion de dialogues qui se veulent ouverts et féconds. Le propos est par ailleurs structuré en des sections courtes et bien agencées, entrecoupées d'encadrés qui permettent d'apporter des points de réflexion supplémentaires, ou un historique des notions discutées. Enfin, on saluera tout à la fois l'autonomie et la complémentarité des différents chapitres, qui permettront à cet ouvrage de servir efficacement d'outil de travail, même si l'âpreté de certaines définitions ainsi que leur foisonnement, l'étendue monumentale du sujet, et le caractère parfois limité des exercices, en font un instrument pédagogique certainement difficile d'accès, malgré la collection de manuels académiques dans laquelle il s'inscrit.

Le premier volume de cet ouvrage se compose de trois parties assez inégales, la troisième constituant les deux tiers de l'ouvrage. La première partie (chapitres 2 à 4) complète essentiellement l'introduction, la deuxième (chapitres 5 à 7) définit formellement les unités de la langue considérées comme pertinentes du point de vue de la syntaxe, et la troisième (chapitres 8 à 13) s'attache à la présentation et à la caractérisation des structures syntaxiques, objet principal de ce volume. S'inscrivant dans la collection « Textbooks in Language Science », l'ouvrage propose également, à l'issue de chaque chapitre, une batterie d'exercices avec leurs corrigés et un aperçu des références pertinentes, le texte étant lui-même enrichi de nombreux encadrés qui mettent en perspective le propos ou viennent offrir un complément (voire un

contrepoint) sur tel ou tel point abordé dans le corps principal. Si l'ouvrage ne se veut pas un manuel de syntaxe en dépendance, les chapitres 9, 10 et 12 constituent néanmoins un ensemble cohérent et très complet qui fournira une introduction tout à fait solide au sujet.

Le chapitre 1, qui fait office d'introduction, situe clairement les objectifs, la portée, et les orientations théoriques de l'ouvrage. Les deux auteurs indiquent ainsi s'inspirer, dans leur démarche, des travaux de Nicolas Bourbaki, un groupe de mathématiciens ayant entrepris de proposer une construction des mathématiques depuis leurs fondations, sans se référer au moindre prérequis et à la moindre notion préexistante. C'est donc l'ambition que se donne cet ouvrage : aboutir à un compte rendu des relations syntaxiques entre éléments de la langue, sans rien considérer comme acquis ou théoriquement établi. Par ailleurs, les auteurs précisent ne pas souhaiter présenter la syntaxe des dépendances comme un outil théorique déjà constitué dont il reviendrait à un manuel d'enseigner le maniement, mais comme un modèle scientifique visant à capturer certaines propriétés des énoncés de la langue, modèle dont on peut dès lors discuter la structure, la motivation, la portée, et la validité.

Le chapitre 2 présente quelques généralités sur la langue, en particulier la notion saussurienne de signe, la distinction langue et parole ou compétence et performance, ou encore les deux perspectives de la théorie linguistique, l'analyse (des énoncés produits) et la synthèse (les règles sous-jacentes à l'expression linguistique d'un sens) privilégiée par les auteurs. Le chapitre 3 explicite justement cette démarche et montre comment un sens peut être représenté schématiquement au moyen d'un graphe orienté, dont différents énoncés alternatifs mais équivalents du point de vue sémantique peuvent alors rendre compte. Cependant, cette perspective orientée vers la production ne sera vraiment développée qu'aux chapitres 12 et 13, et une place très large est laissée à l'analyse des énoncés.

Le chapitre 4 vient donner une assise épistémologique à l'ouvrage en proposant une discussion sur la notion de modèle, laquelle souffre en linguistique d'une certaine imprécision, si bien que la clarification apportée par les auteurs est bienvenue. En prenant pour exemple le modèle gravitationnel des marées en sciences physiques[1], les auteurs mettent en contraste modèle descriptif et modèle explicatif, en soulignant très justement qu'un modèle descriptif a lui-même capacité à prédire s'il est suffisamment développé. Les modèles de type « réseaux de neurones », au fort pouvoir prédictif mais sans valeur explicative, sont également évoqués, les auteurs, s'inscrivant en faux vis-à-vis de ce type, revendiquent de vouloir proposer un modèle proprement explicatif ; à cet égard, on regrettera seulement que la notion d'explication, si difficile en épistémologie, n'ait pas fait elle-même l'objet d'une clarification. Par ailleurs, la dimension prédictive dont se targuent les auteurs n'est en définitive guère sous-jacente dans le reste de l'ouvrage, l'ensemble du modèle

1 Il s'agit au demeurant d'un choix curieux, les marées n'ayant été comprises que tardivement par rapport, par exemple, aux trajectoires planétaires, le modèle newtonien ne parvenant pas à rendre compte des subtilités nombreuses du phénomène sans un effort théorique additionnel.

proposé paraissant essentiellement orienté vers des fins descriptives (ce qui est peut-être un mérite en soi).

La deuxième partie de l'ouvrage, consacrée aux unités de la langue, est introduite par le chapitre 5, qui reprend les principales notions du signe linguistique, et pose une troisième composante en plus du signifié et du signifiant, son syntactique, c'est-à-dire le potentiel combinatoire qui en caractérise l'usage. Dès lors, il se trouve trois manières d'appréhender les constituants minimaux de la langue : du point de vue du sens (sémantèmes), du point de vue de la forme (morphèmes) et du point de vue du syntactique (syntaxème). Cette notion de syntaxème se retrouve justement au cœur du chapitre 6, qui introduit un ensemble de concepts et de définitions fondamentales.

La décomposition propre est d'abord définie : un signe AB peut se décomposer proprement en deux signes A et B, si, par des rapports analogiques du type défini par Saussure, on peut aboutir à un signe A'B' ; par exemple *broyeur* peut se décomposer en *broy* + *eur* parce qu'il existe une analogie *broyeur* : *broyons* égale à *compresseur* : *compressons*. Le morphème est ainsi défini comme un signe qui ne peut pas être décomposé plus avant au sens de la décomposition propre. Cette combinaison propre devient libre lorsque la classe des éléments qui se combinent avec l'un forme une classe indépendante de l'autre. Ainsi, la combinaison de *bless-* et *-ure* n'est pas libre (quand bien même elle est propre) car la classe des éléments qui se combinent avec *-ure* est spécifique de ce signe (avec notamment des radicaux comme *struct-* qui ne correspondent à aucun verbe). Cette combinaison libre permet de définir une unité syntaxique comme un signe se combinant librement avec son environnement. Les unités syntaxiques minimales (qui ne peuvent donc pas être décomposées plus avant au sens de la décomposition libre) définissent alors les syntaxèmes.

Le chapitre 7 complète ces définitions en abordant cette fois les unités sémantiques, définies comme des signes résultant d'un ou plusieurs choix du locuteur. Si ce choix est en plus indivisible, c'est-à-dire qu'il ne peut plus être décomposé en choix successifs, l'unité ainsi délimitée est minimale : il s'agit d'un sémantème. Ce chapitre est également l'occasion de préciser la notion de signème, introduite au chapitre précédent : un signème est défini comme un faisceau de signes, c'est-à-dire un paradigme de formes (comme les différents morphèmes associés au radical du verbe *aller*) mis en relation avec un ensemble de sémantèmes (les différentes acceptions sémantiques). Pour les auteurs, les signèmes (qui sont des constructions théoriques) constituent les unités de la langue, les signes (qui sont des observables), les unités de la parole : la production d'un signe suppose la sélection par le locuteur d'une des acceptions sémantiques du signème, et l'énonciation de la forme appropriée compte tenu des contraintes contextuelles.

La présentation de la syntaxe commence réellement avec le chapitre 8, qui s'appuie sur les chapitres précédents pour préciser la définition des unités syntaxiques (toujours à l'aide de la notion de combinaison libre), qui peuvent être soit minimales (les syntaxèmes), soit une combinaison libre d'unités minimales (les syntagmes). La syntaxe est elle-même définie comme l'étude des combinaisons

libres de signes linguistiques. Le chapitre 9 présente, quant à lui, le concept de connexion syntaxique, qui sous-tend à lui seul la notion de structure syntaxique. Il y a connexion syntaxique lorsque deux éléments syntaxiques se combinent pour former un syntagme – c'est donc, là encore, la combinaison libre qui permet de définir la connexion. Plus précisément, les auteurs définissent la connexion de manière formelle comme une classe d'équivalence, c'est-à-dire un ensemble de combinaisons plus ou moins fines, mais toujours libres, opérant sur la même frontière.

Le chapitre 10 constitue, sans aucun doute, l'apport central de l'ouvrage, tant par son ampleur que par son importance. Il précise en effet la notion de tête et énonce un grand nombre de critères pratiques permettant d'identifier le gouverneur d'une unité syntaxique et, par là, de construire l'arbre de dépendance associé à l'énoncé. On regrettera cependant que la notion de dépendance ne fasse pas l'objet de la même construction formelle que celle, par exemple, de connexion : l'existence d'une hiérarchie entre les unités syntaxiques est immédiatement postulée, sans qu'il soit détaillé ce que signifie la relation de dépendance qui en résulte. Cela n'enlève rien au caractère opératoire de la notion, comme en témoignent les très riches discussions et les nombreux exemples d'applications de critères dont la variété permet de rendre l'analyse en dépendance efficace dans un large éventail de situations. On appréciera également la position très ouverte des auteurs, par exemple sur le choix du nom ou du déterminant comme tête : ceux-ci discutent les différentes possibilités en prenant soin de souligner que la décision finale ne constitue qu'un choix théorique susceptible de révision (ou même d'adaptation à des fins pédagogiques et de présentation).

Le chapitre 11 illustre à nouveau ce positionnement épistémologique qui vise à favoriser le dialogue entre les différentes approches en présentant les relations qu'entretiennent l'analyse en dépendance et l'analyse en constituants. Les limites et les avantages de chacune sont présentés avec détail, et l'exposé, quoiqu'il brasse de nombreux cadres distincts, reste toujours d'une parfaite clarté, supporté notamment par de nombreuses figures qui illustrent bien la fluidité existante entre les divers cadres de représentation. Tout comme le chapitre 10, le chapitre 11 est assorti d'un historique passionnant de ces notions et de ces représentations.

Le chapitre 12 est tout entier dédié à la mise en place d'un modèle topologique, permettant de conformer la structure syntaxique en dépendance à l'ordre linéaire de l'énoncé, dont les contraintes sont propres à chaque langue. Cette approche se veut donc générative dans le sens où ce modèle topologique fournit, à partir d'un arbre en dépendance, les règles permettant la production d'un énoncé linéaire obéissant aux exigences topologiques de la langue. Deux notions sont particulièrement discutées : la projectivité d'abord, réalisée lorsque les dépendances entre unités ne se coupent pas une fois celles-ci arrangées selon l'ordre linéaire, et dont les liens avec la théorie des graphes sont explicités de façon éclairante. La notion de gabarit fait ensuite l'objet de l'essentiel du chapitre : un gabarit est un ensemble de champs ordonnés associé à un constituant, chaque champ venant spécifier quels éléments peuvent l'occuper. Ainsi, pour chaque nœud de l'arbre de dépendance, un gabarit approprié

vient spécifier comment se situent ses dépendants par rapport à lui, suivant la nature de ceux-ci.

Le chapitre 13 se consacre à la structure syntaxique profonde, détaillant le passage d'une structure prédicative du sens à une structure syntaxique profonde, laquelle agence les sémantèmes dans une structure de dépendance, intermédiaire entre le plan du sens et l'arbre de dépendance. Là encore, une approche de type génératif est proposée en associant notamment à chaque unité lexicale un tableau de régime (ou une structure élémentaire d'arbre syntaxique), c'est-à-dire la spécification complète de ses arguments et de leur régime. En combinant les lexèmes suivant cette structure prédicative, on obtient l'arbre correspondant au sens initial. Les auteurs achèvent ainsi de construire l'édifice menant du sens jusqu'à l'énoncé linéairement ordonné.

Silviu PAUN, Ron ARTSTEIN, Massimo POESIO. Statistical Methods for Annotation Analysis. *Morgan & Claypool Publishers*. 2022. 198 pages. ISBN : 978-1-63639-253-0.

Lu par **Lydia-Mai HO-DAC**

Université de Toulouse Jean-Jaurès / CLLE UMR 5263

Cet ouvrage propose un état de l'art « appliqué » des méthodes développées pour évaluer la qualité d'une ressource annotée au niveau linguistique en vue de son utilisation pour le TAL. L'ouvrage se veut avant tout didactique, offrant des explications très claires sur les notions et les formules nécessaires pour maîtriser des méthodes généralement développées dans d'autres domaines que celui de la linguistique (par exemple, la médecine), et des cas d'usage dans le domaine de la linguistique. Les auteurs profitent de chaque explication et cas d'usage pour souligner les précautions à prendre lors de la mise en place de campagne d'annotation et l'importance de l'évaluation dans la constitution de ressources annotées. Trois familles de méthodes sont présentées, des plus simples aux plus complexes (et récentes) : mesures traditionnelles de l'accord entre annotateurs, modèles probabilistes sur l'accord, modèles probabilistes sur l'annotation. La diversité des cas d'usage proposés pour illustrer chaque méthode a été pensée pour couvrir la réalité de la diversité des annotations en linguistique : annotation par les foules vs *par quelques individus au long court, catégorisation binaire* vs *multicatégorisation, avec ou sans jeu d'étiquettes prédéfini, avec ou sans l'étape de délimitation des unités à annoter, annotation pour évaluer* vs *pour entraîner des modèles.*

Organisation générale de l'ouvrage

Après une introduction revenant sur l'activité d'annotation et la différence fondamentale pour les auteurs entre **fiabilité des annotations** (c'est-à-dire le fait que les annotations produites semblent cohérentes entre elles – reliability en anglais) et **validité des annotations** (c'est-à-dire le fait que les annotations soient correctes par rapport à une référence – *validity* en anglais), les auteurs listent certains éléments essentiels à maîtriser pour appréhender l'ouvrage (la notion de processus génératif et d'inférence, l'importance des méthodes existantes pour évaluer l'adéquation entre un modèle et des données, l'influence des préalables dans les modèles statistiques et

les moyens linguistiques à utiliser pour exprimer des probabilités). L'introduction s'achève avec un avertissement sur les connaissances statistiques nécessaires pour une bonne compréhension de l'ouvrage et pointe vers des suggestions de lecture à faire pour acquérir ces connaissances.

Le corps de l'ouvrage est organisé en deux parties qui répartissent les méthodes présentées selon qu'elles mesurent la fiabilité des annotations ou leur validité. Pour chaque famille plusieurs mesures sont présentées, détaillées et illustrées. Après la présentation théorique des concepts de base et des formules impliqués dans la méthode, celle-ci est appliquée à un cas d'usage dont les données sont systématiquement décrites et rendues accessibles sur un site associé (voir *infra*). Des encarts au fil du texte proposent en complément des définitions ou des rappels des notions élémentaires nécessaires pour bien appréhender la ou les mesures présentées (par exemple, l'analyse de contenu, les différentes lois de probabilité utilisées dans le chapitre en cours). Un résumé et un bilan closent la présentation de chaque famille de méthodes permettant ainsi de dégager les spécificités, les avantages et les inconvénients des différentes mesures.

Tout au long de l'ouvrage, des sections sont proposées pour alerter sur les biais possibles auxquels il faut s'attendre au moment de l'évaluation de la fiabilité ou de la validité des annotations.

Un site est proposé en complément de l'ouvrage pour en accompagner la lecture et donner accès aux données mentionnées dans les différents cas d'usage utilisés pour illustrer les différentes méthodes.

Fiabilité des annotations selon la mesure de l'accord entre annotateurs

Dans cette partie, les auteurs distinguent les méthodes utilisant des coefficients d'accord entre annotateurs (*coefficient of agreement*) des méthodes fondées sur des modèles probabilistes de l'accord (*probabilistic models of agreement*).

Selon la même progression que pour toutes les autres sous-parties, les coefficients d'accord entre annotateurs sont présentés du plus simple au plus complexe, en ce sens que les plus complexes permettent de prendre en compte une plus grande diversité de paramètres. La première mesure présentée est celle du pourcentage d'accord (c'est-à-dire le pourcentage d'items sur lesquels les annotateurs sont d'accord par rapport au nombre total d'items annotés) que les auteurs utilisent pour illustrer à quel point les mesures les plus simples ne sont pas adaptées à la réalité d'une grande partie des annotations produites en linguistique car, entre autres, les catégories à annoter sont rarement distribuées équitablement dans la langue (sans parler de la question de l'unité à annoter qui sera traitée plus loin). Le fait que certaines catégories sont plus fréquemment rencontrées dans la réalité de la langue va avoir une influence sur le comportement des annotateurs qui vont avoir tendance à annoter ces catégories plus facilement et donc plus souvent et/ou avec un plus grand accord entre annotateurs. Il est donc nécessaire de disposer de méthodes qui s'adaptent à la diversité des annotations pour ne pas biaiser les résultats de l'évaluation. Complexifiant au fur et à mesure le type d'annotation à évaluer, les auteurs progressent vers des mesures permettant de manipuler des

paramètres de plus en plus variés : nombre des annotateurs et nécessité de pondérer certaines annotations liées à des catégories plus fréquentes ou faciles à annoter que d'autres.

Les mesures détaillées sont les suivantes : pourcentage d'accord, mesures d'association, coefficient de corrélation, le S de Bennet *et al.* (1954), le π de Scott (1955), le κ de Cohen (1960), le κ de Fleiss (1971), l'α de Krippendorff (1980) et la variante pondérée du κ de Cohen (1968). Toutes ces mesures sont présentées de façon théorique puis appliquées à un même cas d'usage : la classification par deux ou plus d'annotateurs des actes de paroles selon le modèle DAMSL (*Dialog Act Markup in Several Layers*, Core *et al.* 1997). Ce cas d'usage commun permet de généraliser les différences principales entre les coefficients proposés.

Le chapitre 2 s'achève par la présentation des intérêts et des limites de ces méthodes mesurant le coefficient d'accord entre annotateurs. L'intérêt principal est de prendre en compte la part de hasard qui peut intervenir dans une tâche d'annotation. Cet intérêt est explicite dans les termes anglais utilisés pour les désigner : *Chance-corrected Agreement Coefficients*. Mais le hasard est loin d'être le biais le plus important à éviter lorsque l'on évalue un jeu d'annotations. Parmi les limites des coefficients présentés, les auteurs reviennent en détail sur (i) les difficultés pour gérer les « données manquantes » c'est-à-dire le fait que les items associés à des catégories plus difficiles à annoter ont tendance à être moins (bien) annotés ; (ii) la nécessité de prendre en compte les spécificités de certains annotateurs ou d'évaluer des annotations réalisées par les foules ; (iii) l'impossibilité d'évaluer correctement des annotations incluant une étape de délimitation des unités à annoter (à ce stade, seule la catégorisation d'unités présegmentées a été discutée) ; (iv) les problèmes liés à des modèles d'annotation impliquant des catégories présentant une très forte disparité de fréquence d'apparition ou une certaine difficulté à être identifiée.

Le chapitre 3 prolonge le chapitre 2 en illustrant comment les coefficients d'accord entre annotateurs sont utilisés pour évaluer les tâches propres au TAL : étiquetage morphosyntaxique, identification des actes de paroles, reconnaissance des entités nommées, détection de la subjectivité, segmentation thématique, annotation de la prosodie, annotation des phénomènes d'anaphore et de deixis discursive, résumé automatique, désambiguïsation lexicale. Face à cette diversité et cette complexité des annotations linguistiques, force est de constater que les coefficients présentés ne semblent pas les mieux adaptés pour évaluer la qualité d'une ressource annotée linguistiquement. Ce constat permet aux auteurs d'introduire les modèles probabilistes comme une solution préférée pour évaluer la fiabilité des annotations.

Fiabilité des annotations selon les modèles probabilistes

Le principe de base de l'évaluation de la fiabilité des annotations avec des modèles probabilistes consiste à distinguer deux étapes dans l'évaluation : dans un premier temps les items à annoter sont caractérisés selon leur probabilité à être difficile à annoter et, dans un deuxième temps, selon leur probabilité à faire consensus entre les annotateurs. La première étape part du constat général fait par

tout chercheur ayant mené une campagne d'annotation en linguistique : il existe des cas clairs mais aussi des cas limites beaucoup plus difficiles à juger.

Les modèles probabilistes permettant d'évaluer cette difficulté de jugement (et donc le risque d'annoter différemment les uns des autres) sont largement utilisés dans le domaine médical, ce qui explique que les cas d'usage utilisés dans ce chapitre sont issus de ce domaine. Trois premiers modèles sont présentés : les modèles de Aickins's α (1990), de Gwet's AC_1 (2008) et de Guggenmoos-Holzmann (1996). Le dernier diffère un peu des autres en ce sens qu'il cherche à évaluer l'instabilité plus que la difficulté des items (plus un jugement est instable plus il sera dur à faire).

Là encore, les auteurs soulignent les limites de ces modèles, et notamment le fait qu'ils ne permettent pas d'évaluer la validité (voire la véracité) des annotations. Par exemple, ils ne permettent pas d'analyser si les annotateurs font les mêmes erreurs d'annotation (et sont donc d'accord mais sur de mauvaises catégories ou délimitation d'unités). Ils ne permettent pas non plus d'analyser si les annotateurs sont plus enclins à annoter d'une certaine manière. De façon générale, ces modèles ne permettent pas de prendre en compte le biais lié aux aptitudes propres à chaque annotateur, aptitudes qui sont souvent surestimées par les modèles. De plus, ces modèles ne fournissent qu'un score d'accord, sans donner une idée de la qualité de l'annotation en termes de « vérité », ce qui est problématique si l'objectif est d'utiliser ces annotations pour entraîner ou évaluer des modèles.

La solution à ces limites réside, selon les auteurs, dans l'application de modèles de variables latentes (*Latent class model*), ce qu'ils démontrent en appliquant ce type de modèle à deux cas d'usage : annotations en médecine réalisées par un panel diversifié *vs* expert d'annotateurs (Uebersax *et al.* 1989), et classification de phrases en termes de contenu subjectif *vs* objectif où il est plus facile et fréquent d'annoter les items objectifs que les autres (Wiebe *et al.* 1999). Cette dernière partie permet d'annoncer la deuxième partie qui propose d'appliquer le modèle de variables latentes à l'évaluation de la validité plutôt que de la fiabilité.

Validité des annotations selon les modèles de variables latentes

La deuxième partie commence par souligner qu'à l'heure où de plus en plus d'annotations sont réalisées par des foules (*crowdsouring* ou myriadisation) ou des machines (avec les techniques d'*active learning*) plutôt que par des experts, le fait de concentrer l'évaluation sur l'accord entre annotateurs est à revoir ; le plus important étant davantage d'évaluer si les annotations sont correctes ou pas.

Après une introduction sur les modèles probabilistes utilisés pour évaluer la validité d'un jeu d'annotations, les auteurs soulignent un des intérêts majeurs du modèle de variables latentes : la capacité à modéliser à la fois le comportement des annotateurs et l'impact de la difficulté des items sur ce comportement. Concernant le comportement des annotateurs, l'intérêt est de considérer le fait que les annotateurs n'ont pas tous les mêmes capacités d'annotation et qu'ils ne sont pas toujours réguliers notamment à cause de l'influence des items précédemment annotés. Deuxième problème : ces variations de comportement sont augmentées par le degré

et le type de difficulté des items. En effet, certaines difficultés altèrent les aptitudes des annotateurs comme le fait qu'ils appartiennent à des catégories rarement rencontrées, qu'ils soient ambigus (certains items peuvent, selon l'interprétation de l'annotateur, appartenir à des catégories différentes), que les catégories du modèle soient difficiles à distinguer, ou encore que la tâche d'annotation soit lucrative, ou peu motivante ou encore très ludique (par exemple, les *games with purpose*).

Les modèles présentés dans cette partie sont les suivants : le modèle de David et Skene (1979) et celui de Carpenter (2008) pour modéliser le comportement des annotateurs, Whitehill *et al.* (2009) pour la difficulté d'items et trois modèles dont celui défendu par les auteurs pour tenter de prendre en compte ces deux aspects (Simpson *et al.* 2011, Felt *et al.* 2015 et les auteurs Paun *et al.*, 2018). Une dernière option est présentée, le modèle discriminatoire de Raykar *et al.* (2010). Celui-ci propose de caractériser dans un premier temps les items afin de nuancer l'évaluation des annotations par rapport à ces caractéristiques.

Le premier cas d'usage utilisé pour illustrer ces modèles et comparer leurs résultats est la tâche de reconnaissance de l'inférence textuelle où il est demandé aux annotateurs de dire si un argument peut être inféré depuis un premier argument, oui ou non. Les annotations utilisées pour ces illustrations sont issues du travail de Snow *et al.* (2018) qui a récolté des annotations *via* le service d'Amazon Mechanical Turk. Ce cas d'usage est un bon exemple pour illustrer les variations entre annotateurs et les degrés et types de difficulté qu'un annotateur peut rencontrer.

Deux autres cas d'usage permettent une description plus approfondie de l'application de certains des modèles présentés : l'analyse de séquences et l'annotation des anaphores, qui illustrent des cas d'annotation plus complexes car ne faisant pas appel à une catégorisation binaire comme la tâche de reconnaissance de l'inférence textuelle.

L'ouvrage s'achève avec un dernier chapitre qui prône l'intérêt d'utiliser les résultats des évaluations issus des différents modèles pour mieux définir le phénomène linguistique que l'on cherche à annoter et améliorer les schémas et les processus d'annotation afin d'obtenir des jeux de données dont la qualité assurera une certaine validité pour les systèmes d'apprentissage automatique.

L'ouvrage de Paun, Artstein et Poesio est complet et très dense. Chaque formule est expliquée, puis appliquée pas à pas aux cas d'usage qui ont le mérite d'être en lien avec la réalité des annotations linguistiques. Les auteurs montrent bien qu'évaluer un jeu d'annotations est complexe et que de nombreux biais sont présents dans les données, notamment ceux liés au comportement des annotateurs et aux difficultés des items en langage naturel. Enfin, les auteurs discutent régulièrement de l'importance d'évaluer pour s'assurer de la qualité des données sur lesquelles entraîner et évaluer des modèles.

Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Karën FORT : karen.fort@loria.fr

Titre : Myriadisation et éthique pour le traitement automatique des langues

Mots-clés : traitement automatique des langues, myriadisation, éthique, jeux ayant un but.

Title: *Crowdsourcing and Ethics for Natural Language Processing*

Keywords: *natural language processing, crowdsourcing, ethics, games with a purpose.*

Habilitation à diriger des recherches en informatique, LORIA, UMR 7503, sous la direction de Mme Claire Gardent (DR, CNRS). Habilitation soutenue le 23/11/2022.

Jury : Mme Claire Gardent (DR, CNRS, directrice), M. Jean-Yves Antoine (Pr, Université François Rabelais, Tours, rapporteur), Mme Iris Eshkol-Taravella (Pr, Université Paris Nanterre, rapporteuse), M. Dirk Hovy (associate professor, Bocconi University, Milan, Italie, rapporteur), M. Philippe Blache (DR, CNRS, président), Mme Armelle Brun (Pr, Université de Lorraine, examinatrice), M. Massimo Poesio (Pr, Université Queen Mary, Londres, Royaume-Uni, examinateur), Mme Marta Severo (Pr, Université Paris Nanterre, examinatrice).

Résumé : *Le traitement automatique des langues (TAL) a subi deux révolutions ces dix dernières années : le raccourcissement extrême de la distance entre les productions de la recherche et l'utilisateur final et l'avènement de l'apprentissage profond (*deep learning*). En conséquence, les besoins en données ont explosé en parallèle des questions éthiques. Cette habilitation à diriger des recherches présente les travaux que j'ai menés dans le domaine de la production d'annotations manuelles pour le TAL par myriadisation (*crowdsourcing*), en particulier par le jeu (*games with a purpose*),*

et dans celui de l'éthique pour le TAL. J'y redéfinis la myriadisation et les sciences participatives en général et je présente en détail les jeux ayant un but, leurs atouts et leurs limites. Je m'attarde plus particulièrement sur ZombiLingo, qui a servi à collecter des annotations en syntaxe de dépendances pour le français et RigorMortis, un jeu d'annotation d'unités polylexicales. Je me concentre dans une dernière partie sur l'éthique pour le TAL, un sous-domaine qui n'a véritablement été reconnu qu'à partir de 2016 et dont j'ai été précurseure. Je reviens sur son historique, son évolution récente et présente mes travaux, menés dans une optique plus déontologiste que conséquentialiste, permettant d'avoir une vision systémique du TAL et des problèmes éthiques qu'il pose.

URL où le mémoire peut être téléchargé :
https://theses.hal.science/tel-03873000

Jade MEKKI : jade.mekki@gmail.com

Titre : Caractérisation de registres de langue par extraction de motifs séquentiels émergents

Mots-clés : registres de langues, traitement automatique des langues, motifs séquentiels.

Title: *Characterisation of Language Registers Using Emerging Sequential Pattern Extraction*

Keywords: *language registers, natural language processing, sequential patterns.*

Thèse de doctorat en informatique, Expression, IRISA, UMR 6074, Université de Rennes 1, sous la direction de M. Damien Lolive (MC HDR, Université de Rennes 1), Mme Delphine Battistelli (Pr, Université de Paris Nanterre), M. Gwénolé Lecorvé (chercheur, Orange) et M. Nicolas Béchet (MC, Université de Bretagne-Sud). Thèse soutenue le 08/09/2022.

Jury : M. Damien Lolive (MC HDR, Université de Rennes 1, codirecteur), Mme Delphine Battistelli (Pr, Université de Paris Nanterre, codirectrice), M. Gwénolé Lecorvé (chercheur, Orange, codirecteur), M. Nicolas Béchet (MC, Université de Bretagne-Sud, codirecteur), Mme Farah Benamara (MC HDR, Université Paul Sabatier, rapporteuse), M. Thierry Charnois (Pr, Université Paris 13 Nord, rapporteur), M. Jean-Yves Antoine (Pr, Université François Rabelais, Tours, président), M. Olivier Baude (Pr, Université de Paris Nanterre, examinateur), M. Dominique Legallois (Pr, Université Sorbonne Nouvelle, examinateur).

Résumé : *Le locuteur d'une langue sent que pour un même message il existe plusieurs manières de le dire. Ce phénomène linguistique est celui des registres de langue. Ils renvoient à un trait saillant du langage que tout locuteur saisit intuitivement.*

Cette thèse s'intéresse à la caractérisation automatique des registres. Notre approche se fonde sur un large corpus de textes, considère divers niveaux d'analyse de la langue en même temps et caractérise les registres de manière comparative.

Sur le plan linguistique, notre contribution est d'étudier les apports des techniques de traitement automatique des langues pour extraire de nouvelles connaissances à propos des registres familier, courant et soutenu. Sur le plan informatique, nous avons proposé une méthode suffisamment générique et non supervisée pour caractériser tout type de variation linguistique, les registres s'apparentant alors à un cas d'usage.

Dans le manuscrit, nous dressons tout d'abord un état des lieux des multiples définitions présentes dans la littérature, par rapport auquel nous positionnons nos travaux. En effet, si les registres de langue semblent être un phénomène intuitivement reconnaissable et aisé à saisir, il n'existe aucun consensus sur leur définition dans la littérature scientifique.

Nous présentons ensuite la constitution linguistiquement motivée d'un large corpus de tweets en français étiquetés en registres. Les étiquettes découlent d'un procédé semi-supervisé fondé sur une graine annotée manuellement en registres et un classifieur qui généralise les annotations à l'ensemble des tweets. Le corpus étiqueté en résultant compte 228 505 tweets pour un total de 6 millions de mots.

À partir de ce corpus étiqueté, nous montrons que l'emploi de techniques d'extraction de motifs séquentiels émergents permet d'extraire des traits linguistiques caractéristiques des registres étudiés. Notre approche lève les trois principaux verrous de la fouille de motifs : la complexité algorithmique, l'abondance des motifs extraits et la difficulté d'évaluer ces derniers.

Le premier verrou est levé avec une méthodologie s'appuyant sur un ensemble minimal de traits linguistiques pour décrire chaque mot mais dont les croisements maximisent la description. La seconde limite est endiguée en réduisant le nombre de motifs extraits en les partitionnant automatiquement. Enfin, le dernier verrou est désamorcé en proposant deux protocoles d'évaluations indépendantes (automatique et perceptuelle).

Le manuscrit s'achève avec l'application de notre approche à une autre problématique : caractériser différents genres de textes adressés aux enfants. Les résultats obtenus indiquent que notre approche est robuste et peut être généralisée à d'autres cas d'usage.

URL où le mémoire peut être téléchargé :
https://hal.science/tel-03991094

Filip MILETIC : filip.miletic@ims.uni-stuttgart.de

Titre : Étude des glissements de sens induits par le contact de langues en anglais québécois : apports conjoints de la modélisation vectorielle sur corpus et de l'approche sociolinguistique variationniste

Mots-clés : glissements de sens, anglais québécois, modèles sémantiques vectoriels, corpus de tweets, sociolinguistique variationniste, contact de langues.

Title: *An Investigation into Contact-Induced Semantic Shifts in Quebec*

Keywords: *semantic shifts, Quebec English, vector space models, Twitter corpora, variationist sociolinguistics, language contact.*

Thèse de doctorat en sciences du langage, CLEE, Université Toulouse - Jean Jaurès, sous la direction de Mme Anne Przewozny-Desriaux (Pr, Université Toulouse - Jean Jaurès) et M. Ludovic Tanguy (MC, Université Toulouse - Jean Jaurès). Thèse soutenue le 20/06/2022.

Jury : Mme Anne Przewozny-Desriaux (Pr, Université Toulouse - Jean Jaurès, codirectrice), M. Ludovic Tanguy (MC, Université Toulouse - Jean Jaurès, codirecteur), M. Stefan Dollinger (Pr, University of British Columbia, Vancouver, Canada, rapporteur), Mme Sabine Schulte im Walde (Pr, Universität Stuttgart, Allemagne, rapporteuse, présidente), M. Kris Heylen (researcher, KU Leuven, Louvain, Belgique, examinateur), Mme Amélie Josselin-Leray (MC, Université Toulouse - Jean Jaurès, examinatrice).

Résumé : *Cette thèse étudie les glissements de sens induits par le contact de langues en anglais québécois, à savoir des mots anglais préexistants utilisés avec un sens différent en raison d'une influence potentielle du français. Ce phénomène sociolinguistique est décrit dans plusieurs études antérieures, mais il reste de nombreuses inconnues quant à sa diffusion, aux contraintes sur ses usages et à la valeur sociale qu'il véhicule. Nous proposons une approche novatrice à l'intersection du traitement automatique des langues et de la sociolinguistique variationniste, afin de fournir une description exhaustive de ce phénomène ainsi que d'évaluer les contributions des approches sur corpus mises en œuvre ici.*

Afin d'effectuer des analyses computationnelles de variation sémantique, nous avons constitué un corpus composé de 78.8 millions de tweets, publiés par 196 000 locuteurs de Montréal, Toronto et Vancouver. Le corpus a été utilisé pour mettre en œuvre différents types de modèles vectoriels, à savoir des représentations computationnelles du sens des mots. Les modèles statiques ont permis d'identifier de nouveaux glissements de sens (en identifiant des différences entre les locuteurs de Montréal par rapport aux deux autres villes), alors que les modèles contextuels ont permis de caractériser plus finement leurs utilisations. Malgré des résultats prometteurs, les analyses qualitatives indiquent que ces méthodes sont limitées par le bruit lié à leurs caractéristiques intrinsèques et à la structure du corpus. Ceci est corroboré par une évaluation quantitative systématique effectuée sur un jeu de données composé de 80 items. Celle-ci a montré

que des résultats comparables à l'état de l'art sur une tâche classique de détection de changement sémantique ne se traduisent pas directement par la capacité pratique à repérer de nouveaux glissements de sens.

Ces approches à grande échelle ont été complétées par des données plus fines recueillies au moyen d'entretiens sociolinguistiques avec 15 locuteurs vivant à Montréal. Nous avons utilisé un protocole sociophonologique classique, garantissant des résultats comparables et fiables, ainsi qu'un nouveau test de perception portant sur l'acceptabilité de 40 glissements de sens attestés dans le corpus de tweets. Les corrélations entre ces variables linguistiques et différents facteurs sociodémographiques, ainsi que les remarques qualitatives sur leur utilisation, indiquent quatre patterns de variation synchronique. Ceux-ci pourraient à leur tour refléter des processus diachroniques. Par ailleurs, la variabilité inter-locuteurs suggère un rôle important des locuteurs bilingues et plus jeunes dans l'utilisation des glissements de sens. Enfin, les scores d'acceptabilité sont faiblement corrélés avec les mesures computationnelles, ce qui suggère que ceux-ci reflètent d'autres dimensions de variation sémantique.

Dans l'ensemble, cette thèse a fourni la première description systématique, menée sur corpus et au moyen d'entretiens, des glissements de sens en anglais québécois induits par le contact avec le français. Elle a également mis en évidence la complémentarité des approches développées dans des disciplines différentes : notre objet d'étude sociolinguistique a orienté la mise en place des expériences computationnelles ; celles-ci ont fourni les stimuli utilisés dans les entretiens sociolinguistiques ; ces derniers ont apporté une évaluation supplémentaire des méthodes computationnelles. Ces considérations ouvrent la voie à une utilisation plus avisée des méthodes computationnelles basées sur corpus dans des études de phénomènes sociolinguistiques.

URL où le mémoire peut être téléchargé :
https://dante.univ-tlse2.fr/s/fr/item/32205