Using Dependency Syntax-Based Methods for Automatic Detection of Psychiatric Comorbidities

Yannis Haralambous, Christophe Lemey, Philippe Lenca, Romain Billot, Deok-Hee Kim-Dufor

▶ **To cite this version:**

Yannis Haralambous, Christophe Lemey, Philippe Lenca, Romain Billot, Deok-Hee Kim-Dufor. Using Dependency Syntax-Based Methods for Automatic Detection of Psychiatric Comorbidities. Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments, May 2020, Marseille, France. pp.142-150. hal-02861753

HAL Id: hal-02861753

https://imt-atlantique.hal.science/hal-02861753

Submitted on 17 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Using Dependency Syntax-Based Methods for Automatic Detection of Psychiatric Comorbidities

Yannis Haralambous[1], Christophe Lemey[2], Philippe Lenca[1], Romain Billot[1] & Deok-Hee Kim-Dufor[2]
[1] IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238, Brest, France
[2] Adolescents and Young Adults Mental Health Department, Brest Medical University Hospital, Brest, France
& EA 7479 SPURBO, Université de Bretagne Occidentale, Brest, France

Abstract

This paper presents the early stages of a growing corpus of psychiatric interviews from help seeking patients referred to an early detection and intervention center for psychosis. In order to contribute to the practitioner's diagnostic, we focus on a new method of automatic comorbidity detection in the corpus. Among the novelties of this method is the fact that it is based on syntactic features of paralinguistic data (interjections and pauses). We use the formalism of dependency syntax, a brief description of which we provide in the paper. Considering the (currently) small size of the corpus, our intention is to prove the applicability of the method rather than to obtain general results about the relevance of syntactic indicators.

1. Introduction

According to the 2001 report of the World Health Organization, psychotic disorders (among which schizophrenia) are one of the main public health problems (Anderson, 2019). They are the third disease in terms of disabilities for individuals (Rössler et al., 2005). This chronic and disabling pathology has an important functional and social impact. It may lead to addictive and self-harming behaviors and brings about severe pain in the patients and their relatives. Schizophrenia is a disease that sets in progressively and at various speeds from one individual to another. The symptoms are diverse and unspecific in the stage preceding the prodromal phase (Yung and McGorry, 1996). In addition, these disorders often arise during adolescence which is characterized by upheavals. The evolutive course of schizophrenia is as follows: the premorbid phase, from birth of the patient until the emergence of the first signs; the prodromal phase during which appear scarcely specific first signs of the disease; these unspecific symptoms gradually increase in intensity and specificity during the phase that precedes the clear psychotic symptoms. Eventually, the psychotic phase arises with the known first psychotic signs that determine the onset of psychosis. The active phase of schizophrenia is characterized by a sheaf of very variable symptoms:

1. positive symptoms: delirious ideas and hallucinations;
2. negative symptoms: social withdrawal and cognitive deficits;
3. disorganization syndrome: contact disorder.

About 600,000 people are currently (early 2020) diagnosed with this disease in France[1] and it is notable that one out of two patients attempts to commit suicide during the evolution of the illness (Castelein et al., 2015). Furthermore, marijuana abuse correlates highly with the risk of developing the disease by doubling it (Krebs et al., 2019). It is a complex disease the physiopathology of which remains little known. The current world-wide dominant explanatory model is the diathesis-stress model that combines two factors: intrinsic vulnerability and stress originating in lived experiences (Howes and McCutcheon, 2017; Bernardo et al., 2017; Pruessner et al., 2017; Millman et al., 2018). Nevertheless, the underlying mechanisms still need to be explored.

The duration between the appearance of the first clear psychotic symptoms and the first access to care is on average two to five years when considered on a world-wide level (differences between various regions being quite important). This period is commonly called "duration of untreated psychosis" (DUP) (Fusar-Poli and others, 2013). Efforts head towards an early treatment and a reduction of the DUP. Indeed, the early identification and rapid interventions during the evolution of a psychotic disorder seem to maximize the therapeutic effects and improve the patients' quality of life (McGlashan and Johannessen, 1996). During this phase, warning signs prior to the active phase of the disease can be detected, and this results into optimization of care and reduction of the DUP (Olsen and Rosenbaum, 2006). It is these very symptoms that lead patients to medical centers and draw the attention of medical staff for early detection. The patients with unspecific symptoms hinting at the onset of schizophrenia are referred to specialized consultations at centers for early detection of psychosis, for the sake of a further evaluation of each patient's symptoms. The populations involved are young adults and have previously demonstrated, for the most of them, a suicidal idea or gesture, or behaviors impacting their emotional, social or professional life (Hutton P, 2011). Various studies have resulted into the development of assessment tools (Olsen and Rosenbaum, 2006; Schultze-Lutter, 2009; Yung et al., 2005).

1.1. Language Analysis in Psychiatry

Speech, and therefore language, is one of the key elements that clinicians can draw on during psychiatry consultations in order to better understand the patients' psychological conditions. Psychiatrists are often led to study its phonetic, syntactic and semantic features, which are likely to reveal pathological conditions. Patients with schizophrenia may demonstrate thought disorder, i.e., disorganized

[1] https://www.inserm.fr/information-en-sante/dossiers-information/schizophrenie

thought, which is a characteristic element of this disease. It has been shown that speech analyses can measure thought disorder (Mota et al., 2012). Techniques of computerized speech analyses such as latent semantic analysis, discourse analysis using graph theory and structural discourse analysis have demonstrated a decrease in coherence in patients with schizophrenia correlated with the clinical evaluations and an identical or higher accuracy of diagnosis (Mota et al., 2012; Hoffman et al., 1986; Elvevåg et al., 2007). Through these approaches the first-degree relatives of patients with schizophrenia can be distinguished from control subjects (Elvevåg et al., 2010), and subtly disorganized elements in high-risk patients' speech—which predicts a transition to psychosis—stand out (DeVylder and others, 2014). It has been shown that a combination of semantic and syntactic analyses can predict with reasonable accuracy the transition to schizophrenia and seems to be more efficient than the standard clinical evaluation (SIPS 79%) (Bedi et al., 2015), This method has been replicated in an independent cohort (with an accuracy of 83%) (Corcoran et al., 2018). Prosodic analyses led by different international teams on psychiatric comorbidities have focused mainly on the fundamental frequency (F0) and speech rate (Scherer and Bänziger, 2004; Audibert et al., 2005; van den Broek, 2004; Moore et al., 2003). (Silber-Varod et al., 2016) have in common with our approach the fact that they consider pauses and disfluencies in anxiety comorbidities—nevertheless their work is mainly based on prosodic characteristics, while we focus on syntax.

1.2. Psychosis risk assessment

Within the scope of consultations for early detection and intervention, numerous patients are received and have gained access to further assessment of their disorders. The centers for evaluation of risk for psychosis receive patients addressed to them by health, social and care partners who are often helpless before the emergence of a non-constituted psychiatric disorder which manifests itself through an unspecific and polymorphic symptomatology (Le Galudec et al., 2014).

Patients received in our center for early detection (at the Adolescents and Young Adults Mental Health Department, Brest Medical University Hospital) are assessed by a multidisciplinary team including, but not limited to, a psychiatrist, a psychologist, a nurse and a neuropsychologist (Bazziconi et al., 2017). The initial evaluation allows identification of a risk level and the establishment of a personalized care protocol, adapted to the intensity of the disorders. A biannual reevaluation is proposed for two years in order to identify potential aggravations of the disorders and any possible onset of psychosis. This transition of the status of patients "at risk for developing a psychotic disorder" to the onset of a confirmed psychotic pathology is called "transition to psychosis". Indeed, 24% of the patients at risk develop a psychotic disorder within the following two years and 33% within the following three years (Bazziconi et al., 2017).

It should be noted that the prevalence of comorbidities, especially mood disorder, anxiety and addiction, is very high (Bazziconi et al., 2017). It is important to identify comorbidities so that patients can have access to appropriate care.

1.3. Comorbidities

Comorbidities are evaluated using a standardized clinical interview, i.e., the *Mini International Neuropsychiatric Interview* (MINI) (Sheehan et al., 1998). The following disorders are explored:

A. Major Depressive Disorder, which we subdivide into:
  A_1. Major Depr. Disorder w/o Psychotic disorder
  A_2. Major Depr. Disorder w/ Psychotic features
B. Dysthymia
C. Suicidality
D. (Hypo)manic Episode, which we subdivide into:
  D_1. Hypomania
  D_2. Mania
E. Panic Disorder
F. Agoraphobia
G. Social Phobia
H. Obsessive Compulsive Disorder
I. Post-traumatic Stress Disorder
J. Alcohol Dependence/Abuse
K. Drug Dependence, which we subdivide into:
  K_1. Opioids
  K_2. Cocaine
  K_3. Cannabis
  K_4. Sedatives
L. Psychotic Disorders
M. *Anorexia Nervosa*
N. *Bulimia Nervosa*
O. Generalized Anxiety Disorder
P. Antisocial Personality Disorder

We have grouped the psychiatric comorbidities listed above into three groups depending on the nature of the disorders, in order to make it possible to carry out statistical analyses on a limited number of samples: anxiety disorders (**ANX**) (E, F, G, H, I, O); thymic disorders (**THY**) (A, B, C, D); and addictive disorders (**ADD**) (J, K, M, N). Comorbidities L and P have not been explored in the present study.

1.4. Project Framework

The results presented in this paper enter into the frame of a research project on informal speech analysis involving all of the patients referred to the early detection and intervention center. The research protocol (NCT03525054) was submitted to, and accepted by, the Institutional Review Board (Comité de Protection des Personnes Est-III, N CPP: 18.04.03). It provides a recording of the initial medical clinical interview and a two-year follow-up.

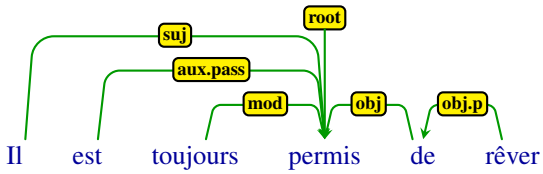
In the following we will first give a short introduction to the specific tool we will be using, namely syntactic dependency relations (§ 2). After that we will describe our corpus and methodology (§ 3) and the results we obtained (§ 4). We conclude with a very short conclusion.

2. Dependency Grammars

In the 19th century two linguists from New York, Alonzo Reed and Brainerd Kellogg, introduced a method for rep-

resenting syntax relations in a graphical way involving only words occurring in the sentence, and thus avoiding the use of word groups. These diagrams were used in schoolbooks starting from 1877, and lasted way into the 20th century. We don't know whether Lucien Tesnière, a French linguist who studied linguistics in Leipzig, Germany, was aware of their existence, but in the late thirties he started working on a new syntactic theory also based on relations between words, which was published after his death (Tesnière, 1959). At that time, linguists were mainly focused on Chomsky's generative transformational grammars, and so Tesnière's work drew almost no attention outside France. And it would probably stay that way, were there not a researcher from Rand Corporation, David Hays, who introduced Tesnière's ideas to the still young community of computational linguists through a presentation at the notorious UCLA symposium on Machine Translation in 1960 (Hays, 1960), a paper in the *Language* journal in 1964 (Hays, 1964) and, finally, in a book that happened to be the first book dedicated to computational linguistics (Hays, 1967). It was he who introduced the terms *dependency grammar* and *dependency relation*. After Hays, the use of dependency grammars continued to spread and nowadays one can reasonably say that they have largely supplanted methods based on constituents in NLP processes (Kübler et al., 2009; Osborne, 2019). Dependency grammars have already been used in the psychiatric domain, for example in (Tanana et al., 2016) where motivational interviewing sessions have been coded via computer.

In a dependency grammar, each sentence has a *head* (usually the verb) that is the root of a directed tree of *dependency relations*. Edges are directed in such a way that one can draw (directed) paths from each leave to the root. Every edge has a tag, called *dependency nature*, which describes the relation between the *dependent* (source of the edge) and the *governor* (target of the edge). Here is a dependency tree example, taken from the *French Treebank Corpus* (Abeillé et al., 2003):



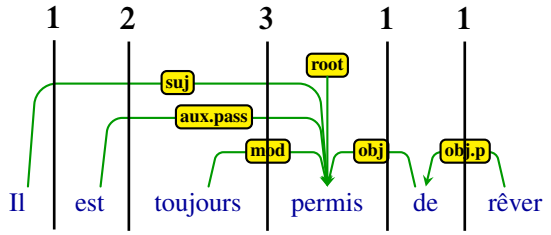
We notice in this example ("It is always allowed to dream") that the participle "permis" is the root of the sentence, and that it governs:

- the pronoun "il" as its subject (suj);
- the verb "est" as its auxiliary verb (aux.pass);
- the adverb "toujours" as a modifier (mod);
- the preposition "de" as its object (obj).

Furthermore, we see that "rêver" is governed by "de" through a prepositional object (obj.p) dependency relation.

2.1. Interstitial Dependency Crossings

Haitao Liu, in (Liu, 2008), explores dependencies from a cognitive point of view and defines a language complexity measure (*MDD = mean dependency distance*) that quantifies the fact that a sentence such as "The man the boy the woman saw heard left," although being grammatical, is more difficult to understand than the equivalent "The woman saw the boy that heard the man that left" (the former has an MDD value of 3 and the later an MDD value of 1.4). By the definition of MDD as the average distance between governor and governed, the more "long-distance" dependencies we have, the higher is the MDD value.
Furthermore, dependencies do not overlap, so that we have an irreflexive, asymmetric and transitive relation $\prec$ between them: $(a \to b) \prec (c \to d)$ when $(\mathrm{pos}(a) < \mathrm{pos}(c)$ and $\mathrm{pos}(d) \geq \mathrm{pos}(b))$ or $(\mathrm{pos}(a) \leq \mathrm{pos}(c)$ and $\mathrm{pos}(d) > \mathrm{pos}(b))$. In the example above, we have (toujours $\to$ permis) $\prec$ (est $\to$ permis) $\prec$ (Il $\to$ permis). The relation $x \prec y$ also implies that $\mathrm{length}(x) < \mathrm{length}(y)$.
The binary relation $\prec$ is a partial order so that we can build a lattice the nodes of which are dependencies and edges represent $\prec$. The lengths of paths in this lattice can be visualized in the dependency tree by drawing vertical lines between words:



Here, the fact that (toujours $\to$ permis) $\prec$ (est $\to$ permis), which is path of length 2 in the lattice, is represented by the fact that the second vertical line crosses two dependencies. Similarly, the the path of length 3 (toujours $\to$ permis) $\prec$ (est $\to$ permis) $\prec$ (Il $\to$ permis) in the lattice, is represented by the fact that the third vertical line crosses three dependencies. As we see, the number of crossings increases when we approach the root from the left since many dependencies targeting the root accumulate, while on the right, because of adjacency between nodes, the crossing number remains low.
Besides the number of crossings, we also use the nature of crossed dependencies in our calculations, e.g., {suj}, {suj,aux.pass}, {suj,aux.pass,mod} on the left and {obj} and {obj.p} on the right, in the example above.

The reason we are interested in interstitial crossings is that in our corpus, besides words we also have *paralinguistic elements*, such as *interjections and pauses*, which occur in interstitial positions.
Our hypothesis is that interstitial positions with a high crossing value are "strategic" and that placing "intruders" (interjections, pauses) in them can be as indicator of some kind of disorder. As we will see, in our small corpus, the combined number and nature of crossed dependencies over a pause or an interjection prove to be comorbidity indicators.

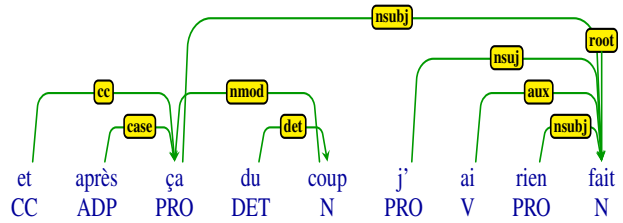
2.2. Parsing Informal Text

One of the major difficulties of this project was the inability of parsers trained on standard language corpora to parse in-

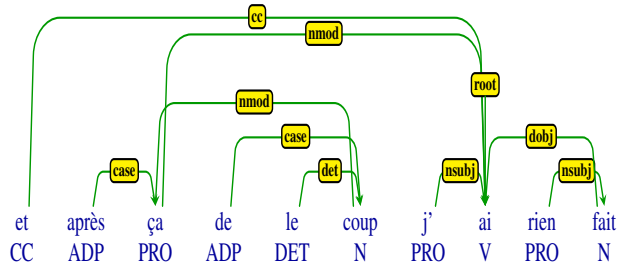
formal text. We will illustrate this by the syntactic analysis of a typical informal French utterance (by patient #44):

et après ça du coup j'ai rien fait

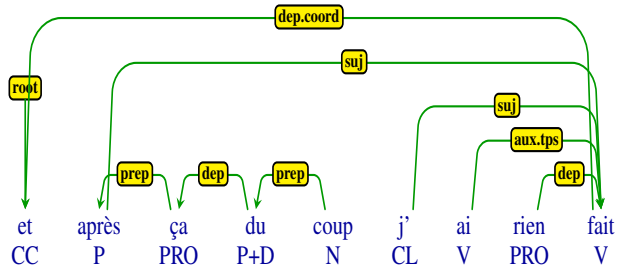
meaning roughly "and after this I haven't done anything". This is a complete turn of the patient, located at approx. the end of the first third of the interview.
Here is the result of the (quite popular) spaCy analyzer (Choi et al., 2015):



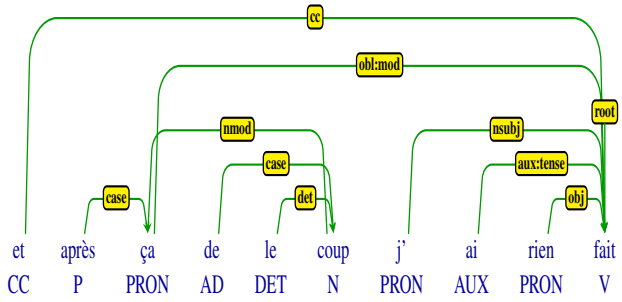
As the reader can see, the word "fait" is taken to be the root of the sentence, but, it carries not a verb tag but a noun tag. Furthermore, the pronoun "rien" is considered as its subject. Also, it is stated that "fait" has a second subject, at the beginning of the sentence, namely the word "ça" (a contraction of "cela"), which is a demonstrative pronoun. It is a strange fact, that "fait" has been chosen to be the root of the sentence but is not tagged as a verb.
The output of the Stanford parser (5/10 2018 version) (Manning et al., 2014) is somehow better:



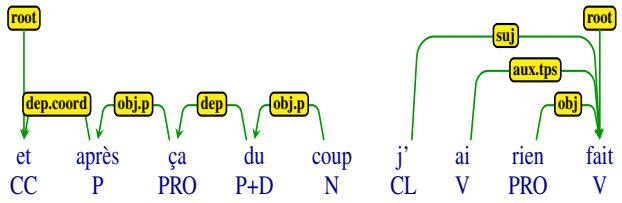
This time it is the auxiliary verb "ai" which is chosen as root of the sentence, and "fait" is its direct object—it is, once again, considered as a noun. Notice that the Stanford parser decomposes "du" into "de le": a preposition and a determinant. The pronoun "ça" is governed by the root, and the nature of its dependency is that of a modifier, although it would be more natural to take the preposition "après" as modifier and to let it govern "ça" as an prepositional object.
The next output comes from the Talismane parser (Urieli, 2013) and is much better in detecting POS tags:



Here it is, surprisingly, the coordination particle "et" which is chosen as root of the utterance, but instead of letting "et" govern the adjacent preposition "après," Talismane connects "et" with the other extremity of the utterance: the participle "fait" (this time, correctly recognized as a verb). Strangely, the preposition "après" is considered as the subject of "fait". Otherwise, the parse is correct, and one can notice a split of the utterance into two groups: in the left group, dependencies are adjacent, while in the right group, they are all governed by the verb "fait," which therefore *acts* like a root, without been *tagged* as a root, probably because Talismane postulates that there can be only one root in an utterance.
The fourth result comes from the very recent package Stanza (with GSD corpus) (Qi et al., 2020). As the reader can see, Stanza avoids all errors made by the other tools we tested: "fait" is detected as being a verb tagged as the root of the sentence, "j'" is its subject and "rien" its object:



But again, similarly to the other tools, Stanza considers that there must be only a single root in a sentence and a single dependency tree, therefore we have long (and not very relevant) dependencies between "et" and "fait" and "ça" and "fait".
The fifth and last result we present is the output of grew (Guillaume and Perrier, 2015), a tool based on the technique of graph rewriting (Bonfante et al., 2018):



Here, the POS tags are the same as in the Talismane example, since the syntax parser grew uses Talismane as its preliminary POS tagger. We notice immediately that grew has indeed recognized the two groups (which Talismane has also noticed but was unable to separate) and has tagged them separately, each one with its own root. Besides the fact that "du coup" could be considered as a secondary interjection and governed by "et," this analysis is by far the most pertinent, and for this reason we have chosen this tool for our project.
Before closing this section we would like to insist on the fact that this comparison of five renowned syntax parsers is de facto unfair since we are using them for something for which they have not been developed, namely for the

Table 1: Comorbidity values of our corpus' patients

ID	Gender	Duration	A_1	A_2	B	C	D_1	D_2	E	F	G	H	I	J	K_1	K_2	K_3	K_4	L	M	N	O	P	**THY**	**ANX**	**ADD**
15	F	$47'29''$	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
21	M	$47'45''$	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	3	0
23	F	$43'50''$	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	1	0
25	M	$30'59''$	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	1	0
27	M	$25'05''$	1	0	0	2	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	3	2	1
28	M	$27'26''$	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	2	2	0
30	M	$63'08''$	0	1	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	1	1
44	M	$43'13''$	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	1	1	2

analysis of transcribed *informal speech utterances*. Possibly some of them would give better results if the utterance had been properly punctuated, but we preferred—as does also (Blanche-Benveniste, 1990)—not to use punctuation since it is not introduced by the patient but by the secretary transcribing the interview.

3. Corpus and Methodology

3.1. The corpus

Our corpus consists of eight patient interviews, of a duration between 25 and 63 minutes. In Table 1 the reader can see characteristics of the patients (gender) and of the interview (duration) as well as the values of standard comorbidities (A-P). As mentioned above, we have grouped comorbidities in three groups, as follows:

THY *Thymic Disorders* (comorbidities A, B, C, D);

ANX *Anxiety Disorders* (comorbidities E, F, G, H, I, O);

ADD *Addictive Disorders* (comorbidities J, K, M, N).

Interviews have been recorded and transcribed by a medical secretary, following conventions for interjections and paralinguistic respiration given in (Bigi, 2015). These transcripts have been proofread and edited by an independent proofreader.

3.2. Methodology

Each recorded interview is segmented into turns between caregiver and patient, and only the latter is kept for further analysis.

Once the transcription has been carefully verified, the two data streams (sound and text) are supplied to SPPAS (Bigi, 2015), which produces a file with timestamped phonemes, words in standard orthography and words in phonemic representation.

But SPPAS does not capture pauses and gaps/lapses. We therefore use Praat (Boersma and Weenink, 2001) on a noise-filtered version of the sound file to detect pauses and gaps/lapses, and then introduce them into the timestamped phoneme and word data. Praat provides us also with energy, pitch, F1 and F2 data, which align with phonemes and words.

In a different data flow (see Fig. 1) we remove interjections from the transcribed text and perform POS tagging on the result using Talismane (Urieli, 2013; Urieli and Tanguy, 2013), followed by dependency parsing using grew (Guillaume and Perrier, 2015). This process provides us with (relatively clean) CoNLL data.

We then align the two data flows (data provided from SPPAS and data in CoNLL form) using the Needleman-Wunsch algorithm as implemented in bioPython. This provides us with a timestamped version of the CoNLL data.

We then use the timestamps of pauses and interjections to study their crossing with syntactic dependencies.

3.3. Definition and Rationale of PIDC and IIDC

Let us consider the dependency syntax forest[2] of a given utterance P. As the reader can see in Fig. 2, in an utterance (taken from patient #44) such as

> les mamans s'inquiètent toujours et les mamans embêtent toujours ce genre-là quoi soyez zen écoutez

words "quoi" and "écoutez" are not connected to the syntax tree of the two coordinated sentences "les mamans s'inquiètent toujours" and "les mamans embêtent toujours ce genre-là" ("quoi" and potentially "ce genre-là" could also be considered as secondary interjections). We therefore do not have a single syntax tree but tree fragments of varying sizes.

The primary interjection "hein" has not been used for the calculation of syntax dependencies, since we have removed it earlier in the process and reintroduced it afterwards. Indeed, we remove all primary interjections in order to obtain dependencies that are closer to the speaker's intention (and to avoid misinterpretation by the syntax parser which has been trained on a corpus without interjections).

The *IIDC* (Interjection Interstitial Dependency Crossings) method consists in re-introducing interjections into the syntax tree by using their timestamps and observing crossings with dependency relations. As the reader can see in Fig. 2, primary interjection "hein" crosses a dependency relation between the noun "mamans" acting as a subject, and the verb "inquiètent," which is the root of the tree fragment. Another interjection (secondary, this time), "quoi," is not crossing any dependency relation since it is located between distinct syntax trees in the forest.

We act similarly for pauses: *PIDC* (Pause Interstitial Dependency Crossings) is the same method applied to pauses (i.e., silences internal to each patient's turn): by their timestamps we align them with the syntax tree fragments and find crossings between them and dependency relations.

Our hypothesis is the following:

[2]We call it a *forest* because of the lack of connectivity, as in the example in § 2.2.

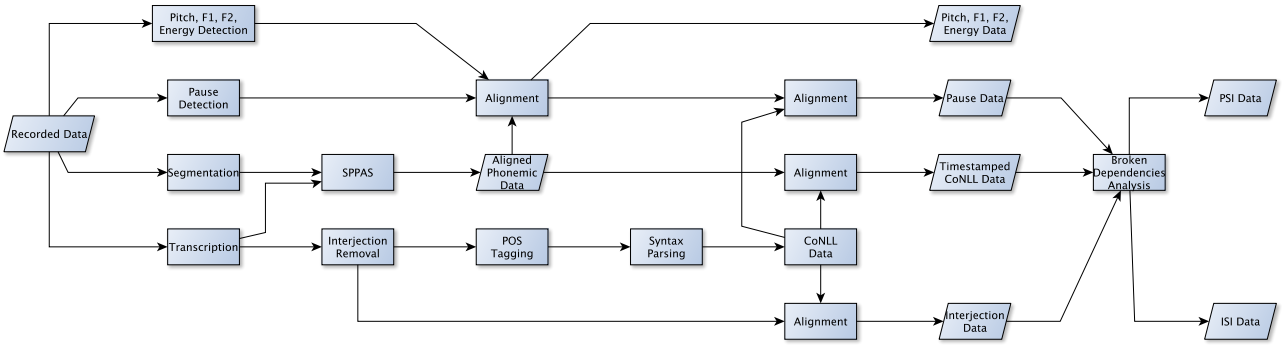


Figure 1: The process of data extraction

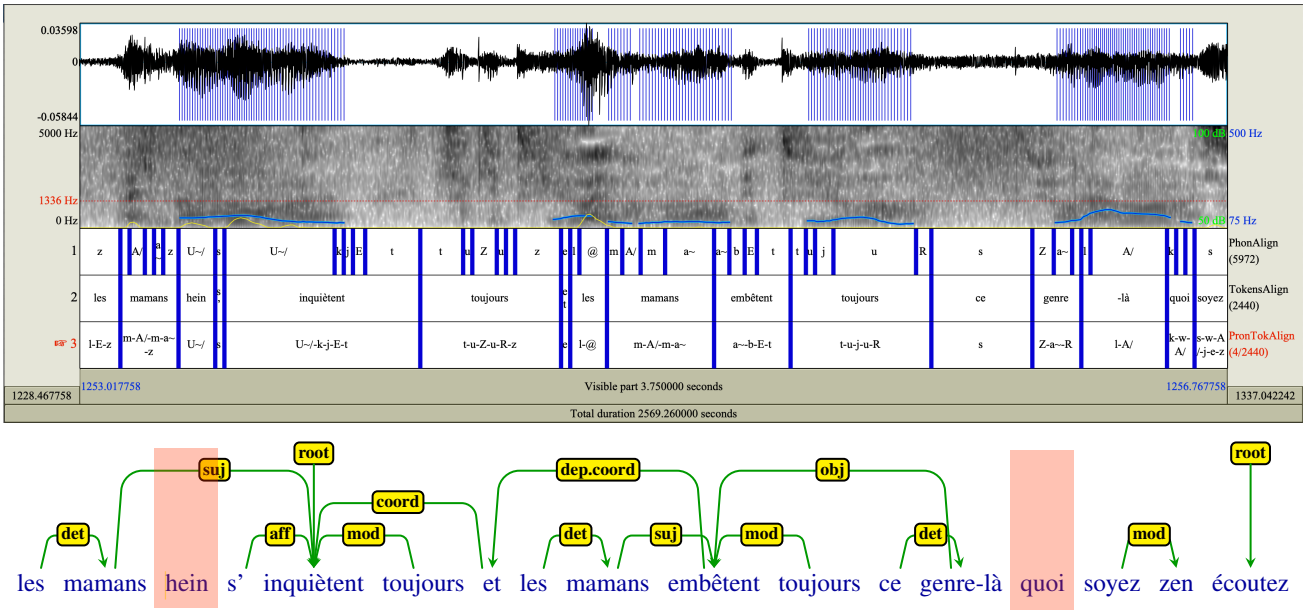


Figure 2: The same utterance (patient #44) visualized in Praat (phonemic alignment) and annotated by syntactic dependencies. The primary interjection "hein" crosses the "suj" dependency "mamans" → "inquiètent". The secondary interjection "quoi" does not cross any dependency relation.

> Interjection Interstitial Dependency Crossings and Pause Interstitial Dependency Crossings can serve as indicators of the patient's linguistic disorganization.

We measure:

1. the number of dependencies crossing interjections or pauses;
2. the nature of these dependency relations.

We define

$$\text{PIDC} := (\#\text{crossings}) \times \frac{\text{pause duration}}{\text{utterance duration}};$$

$$\text{IIDC} := (\#\text{crossings}) \times \frac{\text{interjection duration}}{\text{utterance duration}},$$

and for a given set of dependency relations S we define:

$$\text{PIDC}_S := (\#\text{crossings in } S) \times \frac{\text{pause duration}}{\text{utterance duration}};$$

$$\text{IIDC}_S := (\#\text{crossings in } S) \times \frac{\text{interjection duration}}{\text{utterance duration}}.$$

Besides PIDC and IIDC, we have calculated PIDC_S and IIDC_S for three sets of dependency relations: {det,suj}, {det}, {obj.p} and {suj}. The justification of these choices is as follows.

In the French Treebank Corpus (Abeillé et al., 2003) (which is the most important publicly available dependency-annotated French corpus), among the most frequent relations we note the following:

nature	frequency	avg dist. dep./gov.
mod	120,741	4.1937
obj.p	90,400	1.7511
det	85,154	1.1987
suj	35,402	4.2315

The "mod" (modifier) dependency is very frequent but can take various forms: in 26% of cases the dependent word is an adjective, in 22% of cases a preposition, in 20% of cases an adverb and in 18% of cases a word, and all of these can be located at a certain distance from their governor, therefore the existence of a pause or an interjection between dependent and governor is not necessarily significant.

On the contrary, the "obj.p" (prepositional object) dependency is actually the equivalent of *case government* (for cased languages) and therefore, according to (Osborne, 2019, p. 142), it is technically a type of morphological dependency rather than a syntactic one. It is very stable in terms of POS tag (86% of its dependents are nouns) and the distance between dependent and governor is quite small (1.7511 in average). Its morphological nature and its positional characteristics lead us to formulate the hypothesis that the crossing of an interjection or of a pause with an obj.p dependency is very likely to reveal disorganization.

The "det" (determinant) dependency is also quite suitable to reveal disorganization: the list of determinants is very small and they are very close to their governor (1.1987 in average, that is the smallest average distance above all relations).

Finally the "suj" relation is an important one since (expect in the imperative mode) every French verb necessarily has a subject. We investigate this dependency relation, despite its high distance between dependent and governor (4.2315 in average).

4. Results

We performed a Spearman correlation test on the comorbidity values of the three categories **THY**, **ANX** and **ADD** vs. the various indicators we calculated. Here are the most pertinent results.

We display below the Spearman rho value (and p-value to attest the significance of the results) for comorbidity groups and crossing between pauses/interjections and specific dependency groups:

pause/interj.	{depend.}	group.	rho	p-value
pause	{obj.p}	**ADD**	**0.8660**	0.0054
pause	{det}	**ADD**	0.7735	0.0254
pause	{det,suj}	**ADD**	0.5770	0.1340
pause	{suj}	**ANX**	−0.5086	0.1980
pause	all	**THY**	0.7042	0.0512
interjection	{obj.p}	**ADD**	0.8248	0.0117
interjection	{det}	**ADD**	0.7285	0.04
interjection	{det,suj}	**ANX**	−0.8247	0.0117
interjection	{suj}	**ANX**	**−0.8450**	0.0080
interjection	all	**THY**	−0.6730	0.0671

We notice that for {obj.p} and {det} we get similar behavior for pauses and interjections, even though these two paralinguistic phenomena are quite distinct and have been measured in different ways (pauses have been measured globally by Praat, while interjections have been included by the secretary in the transcription, removed afterwards in order to perform syntax analysis, and re-introduced by their timestamps in SPPAS).

Also we notice that pauses or interjections crossing the {obj.p} dependency are a very strong indicator ($\rho > 0.82$) of the **ADD** group, with a high significance ($p = 0.012$). The {det} dependency also has a consistent behavior (rho around 0.75 with a p-value between 0.025 and 0.04) and, again, targets the **ADD** group.

For the other dependencies, values reveal different behaviors: while pauses crossing {det,suj} or {suj} give insignificant results (p-value > 0.13), interjections combined with {det,suj} and {suj} give very high results, but target negatively the **ANX** group ($\rho < 0.824$ with p-value $= 0.012$). These results can be expressed as follows:

> Members of the **ADD** group tend to place pauses or interjections between preposition and governed noun or between determinant and noun governing it.
>
> Members of the **ANX** group tend to place interjections (but not pauses) between determinant and noun governing it, or between subject and verb governing it.

The first result may reflect the high prevalence of addictive behaviors in patients at risk for psychosis (Valmaggia et al., 2014). As presented previously, the crossing of an interjection or of a pause between preposition and noun or between determinant and noun is very likely to reveal disorganization which is one of the psychotic symptoms often found in at-risk patients (Fusar-Poli and others, 2013). Moreover, the intensity of these psychotic symptoms is correlated with the importance of addictive behaviors (Korver et al., 2010). The second result can be explained by a tendency in anxious patients to avoid leaving gaps, particularly in the context of a conversation where the individual is subject to the judgment of his interlocutor, exactly as would stuttering patients (Iverach and Rapee, 2014).

5. Conclusion

These results show that it is possible to use natural language processing to explore psychiatric comorbidities using linguistic markers. The dependencies and their crossing with pauses and interjection seem to be of particular interest to study in this field. We intend to continue the exploration of linguistic markers following different modalities (semantic, syntactic, prosodic) in order to identify relevant markers for clinical practice.

6. Bibliographical References

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for French. In *Treebanks*, pages 165–187. Kluwer.

American Psychiatric Association APA. (2013). DSM–5: Diagnostic and Statistical Manual of mental disorders, 5th edition. *American Psychiatric Publishing*.

Anderson, K. (2019). Towards a public health approach to psychotic disorders. *Lancet Public Health*, 4(5):e212-3.

Audibert, N., Aubergé, V., and Rilliard, A. (2005). The prosodic dimensions of emotion in speech: the relative weights of parameters. In *European Conference on Speech Communication and Technology*, pages 525–528. ISCA.

Bazziconi, P., Lemey, C., Bleton, L., and Walter, M. (2017). CEVUP program: An analytical epidemiological cohort study. *European Psychiatry*, 41:S729.

Bedi, G., Carrillo, F., Cecchi, G., Fernández-Slezak, D., Sigman, M., Mota, N., Ribeiro, S., Javitt, D., Copelli, M., and Corcoran, C. (2015). Automated analysis of free speech predicts psychosis onset in high–risk youths. *npj Schizophrenia*, 1:15030.

Bernardo, M., Bioque, M., Cabrera, B., Lobo, A., González-Pinto, A., Pina, L., Corripio, I., Sanjuán, J., Mané, A., Castro-Fornieles, J., Vieta, E., Arango, C., Mezquida, G., Gassó, P., Parellada, M., Saiz-Ruiz, J., Cuesta, M. J., Mas, S., and PEPs GROUP. (2017). Modelling gene–environment interaction in first episodes of psychosis. *Schizophr. Res.*, 189:181–189.

Bigi, B. (2015). SPPAS – Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician – International Society of Phonetic Sciences*, 111–112:54–69.

Blanche-Benveniste, C. (1990). *Le français parlé. Études grammaticales*. Éditions du CNRS, Paris.

Boersma, P. and Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *Glot International*, 5(9-10):341–347.

Bonfante, G., Guillaume, B., and Perrier, G. (2018). *Application of Graph Rewriting to Natural Language Processing*, volume 1 of *Logic, Linguistics and Computer Science Set*. ISTE Wiley.

Bowie, C., Gupta, M., and Holshausen, K. (2013). Cognitive remediation therapy for mood disorders: Rationale, early evidence, and future directions. *Can. J. Psychiatry*, 58/6:319–325.

Castelein, S., Liemburg, E., de Lange, J., van Es, F., Visser, E., Aleman, A., Bruggeman, R., and Knegtering, H. (2015). Suicide in Recent Onset Psychosis Revisited: Significant Reduction of Suicide Rate over the Last Two Decades — A Replication Study of a Dutch Incidence Cohort. *PLoS One*, 10:e0129263.

Choi, J. D., Tetreault, J., and Stent, A. (2015). It depends: Dependency parser comparison using a Web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 387–396. Association for Computational Linguistics.

Corcoran, C., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D., Bearden, C., and Cecchi, G. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 17(1):67–75.

DeVylder, J. E. et al. (2014). Symptom trajectories and psychosis onset in a clinical high–risk cohort: the relevance of subthreshold thought disorder. *Schizophr Res*, 159:278–283.

Elvevåg, B., Foltz, P. W., Weinberger, D. R., and Goldberg, T. E. (2007). Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophr. Res*, 93:304–316.

Elvevåg, B., Foltz, P. W., Rosenstein, M., and DeLisi, L. E. (2010). An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *J. Neurolinguistics*, 23:270–284.

Fusar-Poli, P. et al. (2013). The psychosis high-risk state: a comprehensive state-of-the-art review. *JAMA Psychiatry*, 70:107–120.

Guillaume, B. and Perrier, G. (2015). Dependency parsing with graph rewriting. In *International Conference on Parsing Technologies*, pages 30–39.

Hays, D. (1960). Grouping and dependency theories. In *Proceedings of the National Symposium on Machine Translation*, pages 257–266. UCLA.

Hays, D. (1964). Dependency theory: A formalism and some observations. *Language*, 40:159–525.

Hays, D. (1967). *Introduction to computational linguistics*. Macdonald & co.

Hoffman, R. E., Stopek, S., and Andreasen, N. C. (1986). A comparative study of manic vs schizophrenic speech disorganization. *Arch. Gen. Psychiatry*, 43:831–838.

Howes, O. D. and McCutcheon, R. (2017). Inflammation and the neural diathesis-stress hypothesis of schizophrenia: A reconceptualization. *Transl. Psychiatry*, 7:1024.

Hutton P, Bowe S, P. S. F. S. (2011). Prevalence of suicide risk factors in people at ultra-high risk of developing psychosis: a service audit. *Early Interv Psychiatry*, 5, nov.

Iverach, L. and Rapee, R. M. (2014). Social anxiety disorder and stuttering: current status and future directions. *J. Fluency Disord.*, 40:69–82.

Knight, M. J. and Baune, B. T. (2018). Cognitive dysfunction in major depressive disorder. *Curr. Opin. Psychiatry*, 31:26–31.

Korver, N., Nieman, D. H., Becker, H. E., van de Fliert, J. R., Dingemans, P. H., de Haan, L., Spiering, M., Schmitz, N., and Linszen, D. H. (2010). Symptomatology and neuropsychological functioning in cannabis using subjects at ultra–high risk for developing psychosis and healthy controls. *Australian & New Zealand Journal of Psychiatry*, 44(3):230–236.

Krebs, M.-O., Kebir, O., and Jay, T. (2019). Exposure to cannabinoids can lead to persistent cognitive and psychiatric disorders. *European Journal of Pain*, 23.

Kübler, S., McDonald, R., and Nivre, J. (2009). *Dependency Parsing*, volume 2 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool.

Le Galudec, M., Cornily, G., Garlantézec, R., Stéphan, F., Alavi, Z., and Walter, M. (2014). Evaluation of GPs diagnostic knowledge and treatment practice in detection and treatment of early schizophrenia: a French postal survey in brittany. *Social Psychiatry & Psychiatric Epidemiology*, 49:69–77.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9:159–191.

Malhi, G. S., Byrow, Y., Fritz, K., Das, P., Baune, B. T., Porter, R. J., and Outhred, T. (2015). Mood disorders: neurocognitive models. *Bipolar Disorders*, 17(S2):3–20.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

McGlashan, T. H. and Johannessen, J. O. (1996.). Early detection and intervention with schizophrenia: rationale. *Schizophr. Bull.*, 22:201–222.

Millman, Z. B., Pitts, S. C., Thompson, E., Kline, E. R., Demro, C., Weintraub, M. J., DeVylder, J. E., Mittal, V. A., Reeves, G. M., and Schiffman, J. (2018). Perceived social stress and symptom severity among help–

seeking adolescents with versus without clinical high–risk for psychosis. *Schizophr. Res.*, 192:364–370.

Modinos, G. and McGuire, P. (2015). The prodromal phase of psychosis. *Current Opinion in Neurobiology*, 30:100–105.

Moore, E., Clements, M., Peifer, J., and Weisser, L. (2003). Analysis of prosodic variation in speech for clinical depression. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 3, pages 2925–2928.

Mota, N. B., Vasconcelos, N. A. P., Lemos, N., Pieretti, A. C., Kinouchi, O., Cecchi, G. A., Copelli, M., and Ribeiro, S. (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLOS ONE*, 7(4):e34928.

Olsen, K. A. and Rosenbaum, B. (2006). Prospective investigations of the prodromal state of schizophrenia: assessment instruments. *Acta Psychiatr. Scand.*, 113:273–282.

Osborne, T. (2019). *A Dependency Grammar of English*. John Benjamins, Amsterdam/Philadephia.

Pruessner, M., Cullen, A. E., Aas, M., and Walker, E. F. (2017). The neural diathesis–stress model of schizophrenia revisited: An update on recent findings considering illness stage and neurobiological and methodological complexities. *Neurosci. Biobehav. Rev.*, 73:191–218.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*. arXiv:2003.07082.

Rössler, W., Joachim Salize, H., van Os, J., and Riecher-Rössler, A. (2005). Size of burden of schizophrenia and psychotic disorders. *European Neuropsychopharmacology*, 15:399–409.

Scherer, K. R. and Bänziger, T. (2004). Emotional expression in prosody: a review and an agenda for future research. In *The Speech Prosody Conference*.

Schultze-Lutter, F., Ruhrmann, S., Berning, J., Maier, W., and Klosterkötter, J. (2010). Basic symptoms and Ultra-High Risk criteria: Symptom development in the initial prodromal state. *Schizophr Bull*, 36/1:182–191.

Schultze-Lutter, F. (2009). Subjective symptoms of schizophrenia in research and the clinic: The basic symptom concept. *Schizophr. Bull.*, 35:5–8.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., and Dunbar, G. C. (1998). The Mini–International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM–IV and ICD–10. *J. Clin. Psychiatry*, 59 Suppl 20:22–33.

Silber-Varod, V., Kreiner, H., Lovett, R., Levi-Belz, Y., and Amir, N. (2016). Do social anxiety individuals hesitate more? The prosodic profile of hesitation disfluencies in Social Anxiety Disorder individuals. In *Proceedings of the Speech Prosody Conference*, pages 1211–1215.

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., and Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *J Subst Abuse Treat.*, 65:43–50, June. doi:10.1016/j.jsat.2016.01.006.

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

Urieli, A. and Tanguy, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions: études de cas avec l'analyseur Talismane. In *Actes de la 20^e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 188–201.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université de Toulouse II le Mirail.

Valmaggia, L. R., Day, F. L., Jones, C., Bissoli, S., Pugh, C., Hall, D., Bhattacharyya, S., Howes, O., Stone, J., Fusar-Poli, P., and et al. (2014). Cannabis use and transition to psychosis in people at ultra–high risk. *Psychological Medicine*, 44(12):2503–2512.

van den Broek, E. L. (2004). Emotional prosody measurement (EPM): a voice–based evaluation method for psychological therapy effectiveness. *Stud. Health Technol. Inform.*, 103:118–125.

Yung, A. and McGorry, P. (1996). The prodromal phase of first–episode psychosis: Past and current conceptualizations. *Schizophrenia Bulletin*, 22(2):353–370.

Yung, A. R., Yung, A. R., Yuen, H. P., McGorry, P. D., Phillips, L. J., Kelly, D., Dell'olio, M., Francey, S. M., Cosgrave, E. M., Killackey, E., Stanford, C., Godfrey, K., and Buckby, J. (2005). Mapping the onset of psychosis: The comprehensive assessment of at–risk mental states. *Australian & New Zealand Journal of Psychiatry*, 39(11-12):964–971.