MASTER IN COGNITIVE SCIENCE AND LANGUAGE

MASTER THESIS

September 2022

Lexical cohesion and LSA coherence in schizophrenic patients with and without thought disorder

by Pablo del Olmo Encabo

Under the supervision of:

Joana Rosselló Ximenes

Table of contents

Abstract

Language abnormalities called Formal Thought Disorder may be present in patients with schizophrenia. Disorganized speech is one of these abnormalities, and it may be studied by examining the semantic relationships between parts of the text. Two such approaches are lexical cohesion and LSA coherence. Previous studies have found apparently contradictory evidence, with lexical cohesion being increased in the discourse of patients with Formal Thought Disorder and LSA coherence being decreased. In this study, we tested the idea that these findings are not contradictory, and that one may find both these disturbances in the same sample of patients, with both explaining relevant clinical variables independently of each other. Results didn't provide strong support for our hypotheses. We attempt to provide explanations for this failure to find the expected results. Methodological differences with previous literature, low subject number and lack of control of certain symptomatology variables may explain our null findings.

Keywords: Computational Linguistics, schizophrenia, formal thought disorder, discourse, cohesion, coherence

Figures

Tables

List of abbreviations

FTD	Formal Thought Disorder
LSA	Latent Semantic Analysis
TTR	Type Token Ratio
PANSS	Positive And Negative Syndrome Scale
PANSS-P	Positive And Negative Syndrome Scale - Positive
PANSS-N	Positive And Negative Syndrome Scale - Negative
PANSS-GP	Positive And Negative Syndrome Scale - General Psychopathology
GAF	Global Assessment of Functioning
CGI-S	Clinical Global Impression - Severity
TF-IDF	Term frequency–inverse document frequency

Introduction

Schizophrenia is a psychotic disorder characterized by the presence of delusions, hallucinations, disorganized speech and behavior as well as negative symptoms (avolition or flat affect; APA, 2013). Language disturbances occur in some patients with schizophrenia (although they are not unique to this diagnosis) resulting in speech that is, among other characteristics, tangential, lacking in content, incoherent or derailed. These anomalies in language have been called Formal Thought Disorder (FTD; Andreasen, 1986). From a linguistic point of view, FTD has been characterized as an impairment spanning across many language levels, including semantics and pragmatics. It has been noted that FTD concerns expressive language and "higher-order" semantics, understood as the ability to construct larger structures out of individual propositions; as well as coherence, or the "connectedness" of the ideas in discourse (Covington et al., 2005). From a cognitive perspective, FTD has also been linked to semantic deficits (McKenna & Oh, 2005, Chapter 7; Bora et al., 2019). Empirical studies have demonstrated that neuropsychological tests of semantic ability show deficits in FTD patients compared with schizophrenia patients without FTD when these tests are associative ("higher-order" semantic ability) but not when they concern naming (Barrera et al., 2005). Similarly, FTD patients have been shown to be impaired in both the comprehension and repetition production (while preserving gist meaning) of full sentences, indicating an inability to encode and reconstruct linguistic context (Dwyer et al., 2014)

In this study, two apparently contradictory previous findings, both concerning an anomaly in FTD in the use of "higher-order" semantics to produce continuous speech (discourse), will be addressed. On the one hand, Rochester & Martin (1979) found that FTD discourse made *more* use of lexical cohesion (connecting different sentences by means of repeated and semantically related words) than the discourse by schizophrenia patients without FTD. On the other hand, research with LSA and other word embedding models has shown that FTD discourse is characterized by *less* coherence, that is, less semantic similarity between different parts of the text (e.g., Bedi et al., 2015). These findings may seem contradictory, because prima facie, one would expect two measures that both rely on semantic relationships between different parts of the text would vary in the same way across groups. However, an alternative explanation is possible: that these measures are dissociable, capturing different aspects of FTD, both caused by independent disturbances.

Since there have been found dissociations between bottom-up and top-down language processing in schizophrenia (Rabagliati et al., 2019), it is possible that each of these anomalies is due to each of these different aspects of speech production. Although this hypothesis will not be tested here, this study aims to test whether these two anomalies in discourse (increased lexical cohesion and decreased LSA coherence in FTD patients) can be found in the same sample, and whether they independently explain symptomatology and functioning. First, we review findings for each of the disturbances, taking them in turn.

Lexical Cohesion

Haliday and Hasan (1976) developed an annotation method in order to assess the cohesion of normal discourse, that is, the use of linguistic devices to connect independent sentences. They identified five cohesion devices: conjunction, substitution, ellipsis, reference and lexical cohesion. Lexical cohesion was defined as the use of a word or words that were either a complete repetition, a repetition of the root, synonyms, hypernyms, hyponyms or general words that stood for previously uttered words in the same discourse.

Rochester & Martin (1979) applied this annotation method to analyze the speech of schizophrenia patients with and without FTD and healthy controls. Aside of other anomalies regarding less use of overall cohesive ties and reference in the FTD group, they found that, in a free speech task (an unstructured interview) the percentage of lexical cohesion ties out of all ties was higher in the FTD group than in the group of schizophrenia patients without FTD. This result was also found in Rochester et al. (1977). An example of a discourse fragment by a patient with FTD displaying a high amount of lexical cohesion ties can be seen in Figure 1.

Figure 1

An example of thought disordered discourse with a high amount of lexical cohesion

(1) I'm from Marshalltown, Iowa.

(2) That's 60 miles northwest, northeast of Des Moines, Iowa.

(3) And I'm married at the present time.

(4) I'm 36 years old.

(5) My wife is 35.

(6) She lives in Garwin, Iowa.

(7) That's 15 miles southeast of Marshalltown, Iowa.

(8) I'm getting a divorce at the present time.

(9) And I am at presently in a mental institution in Iowa City, Iowa, which is a hundred miles southeast of Marshalltown, Iowa.

Note: (n): sentence number; yellow: examples of repetition, green: examples of semantically related words.

The discourse fragment was obtained from Andreassen (1986), under the section for Perseveration.

Other independent direct replication attempts of this finding are hard to come by, since some authors who did apply Halliday & Hassan (1976)'s cohesion analysis to the study of FTD didn't consider lexical cohesion in thought disordered discourse a true cohesive device (Chaika & Lambert, 1989), focused exclusively on findings concerning the anomalous use of reference in FTD (e.g., Docherty et al., 1988) or used the system to compare between patients with FTD belonging to different diagnostic categories (Wykes & Leff, 1982). However, some studies that have used lexical cohesion to study FTD do exist. Harvey (1983) compared the discourse of schizophrenic and manic patients with and without FTD and healthy controls in an unstructured interview, and he found no significant difference between the FTD group and the non-FTD group (both including patients with schizophrenia and mania) in the amount of lexical cohesion per clause, failing to replicate. However, they did find the expected difference when using the percentage of lexical cohesive ties out of all ties, as Rochester & Martin (1979) had done. Ragin & Oltmanns (1986) compared the discourse of thought

disordered patients with schizophrenia, mania and schizoaffective disorder at admission to an inpatient unit and after discharge, using an interview concerning the events that led to admission. They found that, for FTD patients with mania and schizoaffective disorder, within sentence lexical cohesion was reduced after discharge, with between sentence lexical cohesion remaining stable. For patients with schizophrenia, however, the reduction in the amount of lexical cohesion between sentences after discharge approached significance, with lexical cohesion within sentences remaining stable, providing evidence in line with Rochester & Martin (1979). Finally, in a somewhat related study, Zhang et al. (2021) compared the discourse of patients with post-stroke aphasia to that of healthy controls, and found that, patients with fluent aphasia (which has been recognized as similar in some regards to FTD; Covington, et al., 2005) produced lower proportions of lexical cohesion than controls.

Although direct evidence for lexical cohesion in FTD after Rochester & Martin (1979) is scarce and results are mixed, there are independent lines of research which provide good reason to suppose that, at least in some situations, lexical cohesion will be increased in FTD. Kuperberg et al. (2019) found that masked indirect lexical priming observed as an N400 reduction occurred only in schizophrenia patients and not in controls, and was, in the patient group, associated with decreased activity in the left temporal fusiform area (associated with lexical activation). In a review of the priming literature in schizophrenia, Minzerberg et al. (2002) conclude that when the experimental conditions promote automatic processing (e.g., low stimulus onset asynchronies, masked prime, etc.) increased direct priming effects in schizophrenia can be observed when compared to controls. These “hyperpriming” effects may indicate that increased spread of activation through the semantic network (Collin & Loftus, 1975) or other bottom-up mechanisms producing increased associative activity are a feature of schizophrenia, and may impact language production giving rise, for example, to glossomania or chaining of associations (Covington et al., 2005).

Indeed, perseveration (the inappropriate repetition of responses) is a characteristic of schizophrenia which has been shown to appear during discourse (Crider, 1997) and is a component of FTD (Andreasen, 1986). Some studies have used a Type-Token Ratio (TTR) measure (the ratio of unique words to total words) to attempt to assess the degree of repetition. For example, Manschreck (1981), in Study 1, compared schizophrenia patients with FTD to schizophrenia patients without FTD, patients with affective disorders and healthy controls in a picture description task, and found that TTR was lower in the FTD group than in any of the other three groups. This finding provides evidence that repetition of

words is increased specifically in FTD (although see Crider (1997), who in a review concludes that TTR studies don't have conclusive results).

Summing up this section, it is reasonable to believe that bottom-up uncontrolled associative activity is behind some of the features of FTD, like perseveration or glossomania. Results which show increases of lexical cohesion for thought disordered groups compared to non-thought disordered groups, such as the original finding by Rochester & Martin (1979), may be the product of these disturbances.

LSA Coherence

LSA is a computational model that extracts the meaning of words through the analysis of large amounts of text. It is a distributional semantics model, operating under the assumption that words which are closer together in a text will tend to have similar meanings. The model takes a matrix of term-document co-occurrences as input, reducing its dimensionality through singular value decomposition, and obtaining a matrix in which each word is assigned a high-dimensional vector (a word embedding). Words which are semantically related are assigned vectors with higher cosines between them (Landauer & Dumais, 1997). Through different methods, such as computing the similarity (cosine) between adjacent sentences, LSA has shown to be able to correctly estimate the coherence of texts (manipulated to be more or less coherent), and predict their readability and comprehensibility (Foltz, 2007; Foltz et al., 1998). In the field of schizophrenia research, LSA measures of coherence, or measures of coherence using different but similar word embedding models (such as GloVe and word2vec), have shown to differ between groups with different clinical characteristics. Normally, such measures are fed to a machine learning model, along with other Natural Language Processing or Speech Signal Processing measures, in order to classify participants into clinical and control groups (Voleti, 2019; Hitczenko et al., 2021). In the next paragraphs, we provide a brief review of studies which have used word embedding based measures of coherence to analyze discourse in the speech of psychotic disorders.

Elvevåg et al. (2007), in their experiment 3, compared healthy controls, and schizophrenia or schizoaffective patients with and without FTD in their responses to a structured interview. They plotted the LSA cosine values between the question and a moving window of words and the distance of the window to the question and found the slope of this relationship. They found that greater slopes (greater semantic divergence between the question and the answer as distance increases) was positively related to clinical judgments of tangentiality.

Additionally, with a large window (8 words) they found that patients with FTD produced less coherent answers overall (again, compared to the question) than controls and patients without FTD. Bedi et al. (2015) analyzed free speech in a sample of 34 youths at clinical high risk of psychosis, attempting to predict the incidence of psychotic disorders two and a half years later through the use of a convex hull classifier. Among the linguistic features fed into the classifier (which also included features derived from Piece of Speech Tagging) were various descriptive statistics of a variable obtained by computing the LSA cosine of each sentence either to the next adjacent sentence or to the next to adjacent sentence. The classifier had a 100% accuracy, and lower minimum cosine between adjacent sentences emerged as one of the best predictors of psychosis incidence. Additionally, lower mean and minimum adjacent sentence cosine were among the predictors of subclinical prodromal positive and negative symptoms. The same authors, in Corcoran et al. (2018), attempted to replicate their results in a new sample, although with a slightly different LSA coherence computation (word to word LSA cosines at distances of 5 to 8 words). After factorializing the original variables, they found that high variance, and low minimum and maximum cosines were among the predictors of psychosis in their new sample and in the old one, and had high classification rates in both. No association between the extracted text features and symptomatology was found, but there was one with FTD related human rated scores.

Iter et al. (2018) attempted to replicate the Elvevåg et al. (2007) and Bedi et al. (2015), calling their methods the Tangentiality Model and the Incoherence model, respectively, and failed to find the predicted differences between patients with schizophrenia and controls. They point to the presence of filler words, coherence being biased towards longer sentences and word repetitions to explain the failure to replicate. They address the problems by removing all filler and stop words from the text, employing more recent sentence embedding systems (word2vec and GloVe) alongside LSA and using vector weighing schemes to obtain the sentence embeddings. With these corrections, they were able to find the expected differences with all three systems in the Tangentiality model and with GloVe in the Incoherence model (significant results differed in the weighting scheme used). Just et al. (2019) attempted to in turn replicate Iter et al. (2018)'s results in the German language, including patient groups with and without FTD and not using LSA. They failed to find any differences in the Tangentiality Model. They did find the expected differences in the Incoherence Model using GloVe with one of the weighing schemes (controls exhibiting higher coherence than non-FTD patients, who in turn showed higher coherence than FTD

patients), and coherence scores were inversely correlated with positive symptoms (not negative) and FTD variables. In Just et al. (2020) the authors extended their findings using the coherence metric in which differences were found in the last experiment. The mean differences in coherence weren't significant, but coherence was a negative predictor of FTD across all patients. Low coherence predicted FTD category membership versus control, also while including repetitions as predictor (although it lost significance when variables such as neologisms, referential ambiguity, etc. were introduced), but couldn't classify non-FTD patients and controls. Panicheva & Litvinova (2019) compared patients with schizophrenia and controls' written texts (in Russian) about their day, computing cosines between all words in sliding windows of 3 to 8 words. They found that maximum cosine was higher for controls, and the minimum score occurred later in the sequence for controls, but, against some of the previous results, they found a higher minimum cosine for patients.

Some more recent studies have performed more elaborate computations based on word embedding systems. Kramov (2020) created graphs in which the nodes were clauses in a discourse, and the weights between nodes were weighted cosines between the clauses. An average of the weights of the graph, along with coherence scores obtained through the Incoherence Model were significant predictors of schizophrenia status. Sarzynska-Wawer (2021) used ELMo, a model which generates contextualized word embeddings for each instance of a word through the use of neural networks, and used its representations to classify patients from healthy people with a rate of 80 % success.

As a final aside, it is worth mentioning that many studies have applied word embedding models to the speech in verbal fluency tasks as a proxy of discourse. LSA measures extracted from verbal fluency tasks have been shown to distinguish patients, their healthy relatives and controls (Elvevåg, 2010) or predict ratings of FTD and functioning in older schizophrenia patients (Holshausen et al., 2014).

In sum, despite the methodological heterogeneity of the literature and the variation in results, it seems that, in general, coherence scores obtained through the use of LSA and other word embedding systems are lower for patients with schizophrenia, and lower still for those with FTD. Cosine relationships between vector representations in LSA go beyond mere bottom-up word associations. For instance, Landauer & Dumais (1997) report:

> *Consider the sentence, "The player caught the high fly to left field."* […] *the vector average of the words in this sentence has a cosine of .37 with ball, .31 with baseball,*

> *and .27 with hit, all of which are related to the contextual meaning of fly, but none of which is in the sentence. In contrast, the sentence vector has cosines of .17, .18, and .13 with insect, airplane, and bird.* […] *However,* [the vector to represent fly] *is unlike any of the right words. It has cosines of only .02, .01, and −.02 respectively with ball, baseball, and hit (compared to .69, .53 and .24, respectively with insect, airplane, and bird).* (p. 229)

The example suggests that LSA is able to capture the overall gist meaning of sentences and chunks of text. If this is so, low coherence scores obtained through LSA in FTD patients may represent a break-down of the gist meaning in discourse. It is reasonable to suppose that such a deficit may be caused by disturbances in the top-down mechanisms which guide goal-directed speech.

This study

Rabagliati et al. (2019) found that in people with schizophrenia, during speech processing, there is a dissociation between top-down and bottom-up cues. Specifically, patients were specifically impaired in the use of previous discourse cues to constrain interpretation, while the use of lexical cues, such as the verb use, showed no impairment. If the above discussion of lexical cohesion and LSA coherence is true, and they both capture different aspects of discourse production, dependent on bottom-up and top-down disturbances respectively, then they may both be present at the same time in a sample of patients. On the one hand, bottom-up increased associative activity may boost lexical cohesion in the FTD group, and, on the other hand, top-down deficits in goal-directed speech production may drive down LSA coherence in the FTD group.

In this study the discourse of healthy controls, schizophrenia patients with FTD and schizophrenia patients without FTD across conditions of free and narrative speech will be compared. The aim is to replicate Rochester & Martin (1979)'s results regarding lexical cohesion and to obtain measures of coherence through LSA, in a similar but slightly fashion to previous FTD studies, extending and complementing previous findings. The relationship between these two measures of discourse will be explored. Additionally, a final objective is to investigate whether lexical cohesion and LSA scores can both simultaneously explain variance in symptomatology, severity of illness and functioning. The hypotheses are:

1. Lexical cohesion, at least during free speech, will be higher for the patients with FTD than for the patients without FTD and the controls, replicating the results of Rochester & Martin (1979)

2. LSA coherence will be lower for the patients with FTD than for the patients without FTD and the controls, finding results which are in line with previous findings (e.g., Elvevåg et al., 2007; Bedi et al., 2015)

3. Lexical cohesion will be a significant positive predictor of symptomatology and severity, and a negative predictor of functioning among the patient groups.

> Although to my knowledge there are no previous findings which have tested this hypothesis, the prediction stems naturally from the fact that FTD is itself a symptom of schizophrenia (e.g., APA, 2013; Andreasen, 1986) and is related to both severity and social functioning (Roche et al., 2015). If increased lexical cohesion contributes to FTD, then it should relate in the same way to other variables.

4. LSA coherence will be a significant negative predictor of symptomatology and severity, and a positive predictor of functioning among the patient groups.

> The same reasons given for hypothesis 4 apply here. Additionally, LSA coherence has been shown to predict symptomatology in some studies (Bedi et al., 2015; Just et al., 2019; although not all, see, e.g., Corcoran et al., 2018). Also, at least in one study LSA measures extracted from a verbal fluency task predicted adaptive functioning in older patients with schizophrenia (Holshausen et al., 2014).

Method

Participants

The sample in this study consisted in 3 groups of 10 participants each: 10 healthy controls (HC group), mean age at interview of 44 years, mean WAIS IQ of 107, mean TAP IQ (which provides an estimate of premorbid IQ, Gomar et al., 2011) of 101; 10 patients meeting the criteria of schizophrenia but not formal thought disorder (FTD − group), mean age at interview of 44 years, mean WAIS IQ of 89, mean TAP IQ of 99; 10 patients meeting the criteria for both schizophrenia and formal thought disorder (FTD + group), mean age at interview of 44 years, mean WAIS IQ of 86, mean TAP IQ of 96. Each group of 10

participants was randomly selected out of an original sample of 30 participants per group who had undergone interviews, excluding those who had spoken in Catalan during their interview (see Materials below).

No significant differences between the groups in this study were present in age at interview (p = 1.000) or in TAP IQ (p = .497), but there were significant differences between the groups in WAIS IQ (p = .021), with the FTD + group being lower than the HC group (p = .027) and the FTD − group tending in the same direction (p = .084), according to post-hoc Bonferroni multiple comparisons.

Materials

The interviews to be analyzed in this study were all conducted in the Spanish language. They included three parts, only two of them were analyzed here. During the first one, free speech, participants were instructed to freely discuss the topic of their childhood, with occasional prompts and questions by the interviewer; during the second one, participants underwent the Rorschach test (not included in this study); during the third one, narrative speech, participants were required to recount a fairytale of their choosing, and, if the participant was unable to do so or the amount of speech produced was deemed insufficient by the interviewer, recount a movie or a narrative event or anecdote form their own lives. The interviews were recorded and were accessible to researchers in audiovisual format.

To obtain the LSA coherence scores from the text of the interviews an LSA corpus composed of 404.436 documents extracted from the Spanish Wikipedia was used, with 39.566 different lexical items (after lemmatization). The smoothing function was log-entropy. The imposed number of dimensions k during singular value decomposition was 300. The coherence scores were obtained using the API GallitoStudio 2.0 (Jorge-Botana et al., 2013), accessible through http://www.gallitoapi.net/especifications/index.html , using the Compare Text to Text function.

Positive And Negative Symptoms Scale (PANSS; Kay et al., 1987), Global Assessment of Functioning (GAF; APA, 2010) and Clinical Global Impression - Severity (CGI-S; Guy, 1976) scores, were obtained for each participant in either the FTD − or FTD + groups (one lost value for GAF in the FTD − group), assessing positive and negative symptomatology; psychological, social, and occupational functioning; and overall severity of illness, respectively. Additionally, PANSS-Positive (items: delusions, conceptual disorganization, hallucinatory behavior, excitement, grandiosity, suspiciousness, hostility), PANSS-Negative

(items: blunted affect, emotional withdrawal, poor rapport, passive-apathetic social withdrawal, difficulty in abstract thinking, lack of spontaneity & flow of conversation, stereotyped thinking) and PANSS-General Psychopathology (items: somatic concern, anxiety, guilt feelings, tension, mannerisms & posturing, depression, motor retardation, uncooperativeness, unusual thought content, disorientation, poor attention, lack of judgment & insight, disturbance of volition, poor impulse control, preoccupation, active social avoidance) subscales were calculated

All statistical analyses were performed using the software SPSSInc PASWStatistics18.

Procedure

Each of two researchers transcribed half of the 30 interviews (5 per group each), randomly assigned. Once a researcher had fully transcribed the relevant sections of an interview (the free and narrative speech parts) the text of the interview was divided into numbered sentences according to criteria in Annex I. After this process, in order to preserve blindness to condition during annotation, the two researchers exchanged the transcripts, each annotating the 15 interviews the other had transcribed. Annotation was performed independently for free and narrative speech sections. In both sections, a sentence was only annotated if it was produced by the participant (as opposed to the interviewer) and it was included in a run of three or more consecutive sentences also produced by the participant. In the free speech section, annotation started from the beginning of the text and continued until 50 sentences had been annotated, at which point the rest of the section's transcript was discarded from the analysis (including from the LSA analysis below). In the narrative section, however, annotation started from the beginning and continued until the transcript was exhausted.

The annotation system was based on Haliday & Hassan (1976)'s cohesion analysis and adapted to the Spanish language according to criteria in Annex II. The system consists in identifying parts of speech called cohesive items, which in some specific way relate to previously uttered (or in some rare cases called cataphora, to subsequently uttered) parts of speech, called presupposed items. In the case of this study, cohesive items could only be annotated in the sentences liable for annotation, (see paragraph above) but their corresponding presupposed items could be marked anywhere in the text. All five of Haliday & Hassan (1976)'s types of cohesive relationships were annotated by researchers, but in this study we only analyze lexical cohesion. Lexical cohesive items were marked as such when

they stood in a certain semantic relationship to a previously encountered lexical item (their presupposed item). The semantic relationships in question were:

1) The lexical cohesive item is identical (or has an identical lexeme) to the presupposed item

2) The lexical cohesive item is a synonym or hyponym of the presupposed item.

3) The lexical cohesive item is a hypernym of the presupposed item.

4) The lexical cohesive item is a general word (e.g., "thing") standing for the presupposed item.

5) The lexical cohesive item is otherwise closely semantically related to the presupposed item.

Finally, lexical cohesion in a given section (free or narrative speech) is in this study considered the frequency of lexical cohesive items per annotated sentence:

$$\frac{number\ of\ lexical\ cohesive\ items}{number\ of\ annotated\ sentences}$$

The process of transcription, division in sentences and annotation was supervised by a third senior researcher, and all disagreements were resolved by discussion.

LSA coherence was computed for each section of the interview using multiple text to text comparisons. In LSA (see Introduction) each word is represented by a vector. The vector representing a text (more than one word) is given by the sum of the vectors representing the words which compose the text. Here, each annotated sentence's (see annotation above) vector was compared to the vector representing the rest of the text in the section, (excluding the sentence itself but including all other sentences, non-annotated ones as well) and the cosine between them was obtained. Then, the LSA coherence within a section was computed by averaging between all cosines obtained in this way. That is, if the vectors of all *n* annotated

sentences in a section are given by $\{v_i, v_{i+1}, \ldots, v_n\}$ and the vector of the full text of the section is given by , then LSA coherence in a given section (free or narrative speech) is in this study considered:

$$\frac{\Sigma_{i=1}^{n}\left\{cos\left[\left(u-v_i\right),v_i\right]\right\}}{n}$$

If a sentence couldn't be assigned a vector because none of its words were represented in the LSA space, the sentence was excluded from the analysis.

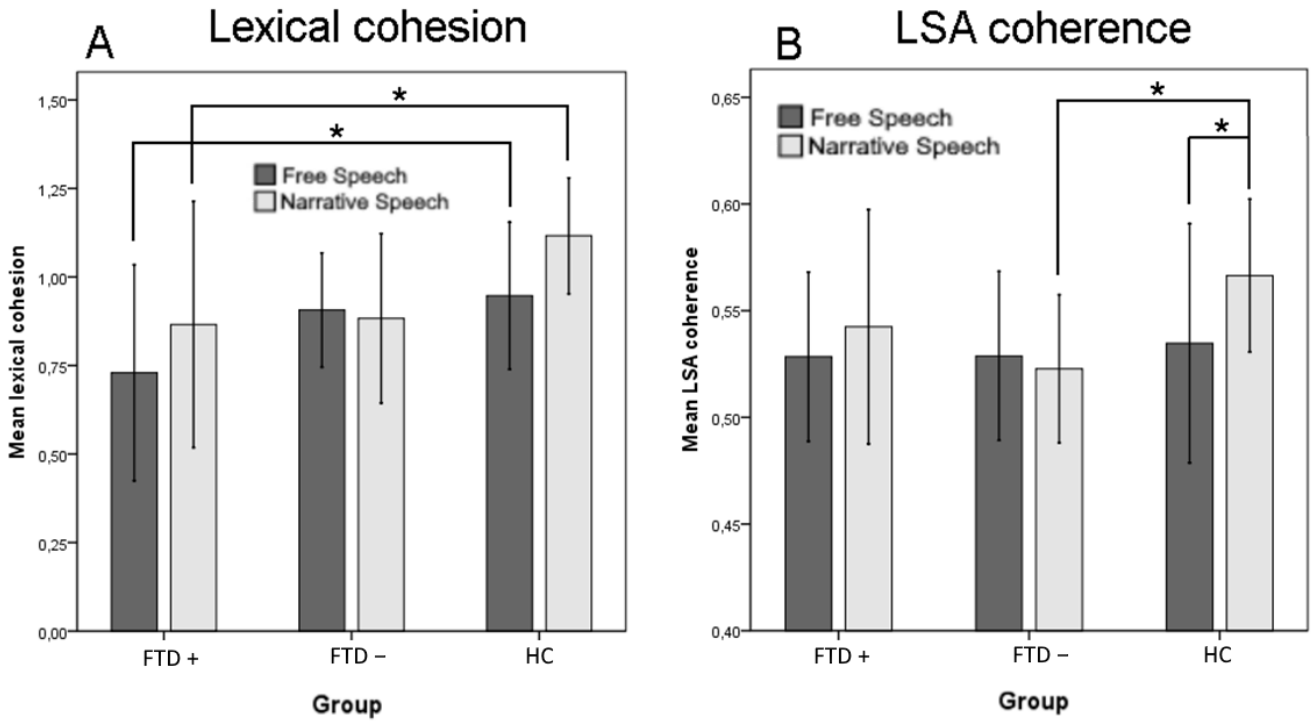
This method to compute LSA coherence, sentence to text comparison, was chosen in order for the results to be comparable with lexical cohesion, in which a cohesive tie may have its presupposed item anywhere else in the text. Although it has to my knowledge not been used in previous schizophrenia or FTD studies (Elvevåg et al., 2007; Bedi et al., 2015) it has shown to be a valid method to measure coherence in a text, distinguishing manipulated coherent from incoherent texts (McNamara et al., 2010).

Results

Firstly, two mixed ANOVAs were run, one with lexical cohesion as a dependent variable and the other with LSA coherence as the dependent variable (see Figure 2 for results). For both, Task (with two conditions: Free Speech and Narrative Speech) was introduced as a within-subjects variable, and Group (with three conditions: HC, FTD − and FTD +) was introduced as a between-subjects variable. Post-hoc pairwise comparisons without adjustment for Type I error were included in both ANOVAs. Comparisons were performed in this way because the number of subjects in each group was low (n = 10) and therefore no further loss in statistical power was desirable. However, Bonferroni corrected *p* values are also reported when relevant

Figure 2

Lexical cohesion and LSA coherence means across levels of Task and Group



Note: Error bars: ± 1 Standard Deviation

* Significance at the .05 level, uncorrected for multiple comparison error

In the ANOVA for lexical cohesion no significant main effect of Task ($F(1, 27) = 2.553$, $p = .122$) or interaction ($F(2, 27) = 1.023$, $p = .373$) were found, but there was a significant main effect of Group ($F(2, 27) = 3.923$, $p < .05$). Pairwise comparisons between the levels of group showed that the only significant difference ($t(2) = 2.786$, $p = .01$) was between the HC group (M = 1,032, SD = 0.112) and the FTD + group (M = 0,798, SD = 0.265), with the HC group showing increased lexical cohesion, and this difference remained significant after Bonferroni correction ($p < .05$). Within levels of Task, there was a significant difference ($t(2) = 2.087$, p

< .05) in the Free Speech condition between HC (M = 0.947, SD = 0.208) and FTD + (M = 0.730, SD = 0.305), and, in the Narrative Speech condition, there was also a significant difference ($t(2) = 2.137$, $p < .05$) between the HC (M = 1.116, SD = 0.164) and FTD + (M = 0.866, SD = 0.348) groups (see Fig. 1A), along with a difference approaching significance ($t(2) = 1.991$, $p = .056$) between the HC and FTD − (M = 0.883, SD = 0.239) groups. However, none of these paired comparisons within levels of Task remained significant or showing a trend towards significance after Bonferroni correction was applied ($p = .138$, $p = .124$, $p = .169$, respectively).

In the ANOVA for LSA coherence, no main effects of Task ($F(1,27) = 2.320$, $p = .139$), Group ($F(2,27) = 1.123$, $p = .340$) or interaction ($F(2,27) = 1,571$, $p = .226$) were found. Despite of this, post-hoc comparisons showed that, in the Narrative Speech condition, coherence in the HC group (M = 0.566, SD = 0.036) was significantly greater ($t(2) = 2.316$, $p < .05$) than in the FTD − group (M = 0.523, SD = 0.035), although the difference does not maintain significance after Bonferroni correction ($p = .092$). Additionally, within the HC group, Narrative Speech showed greater coherence ($t(27) = 2.133$, $p < .05$) than Free Speech (M = 0.535, SD = 0.056), see Fig. 1B. No other pairwise comparisons reached significance.

Secondly, Pearson correlations were calculated between lexical cohesion in Free Speech, lexical cohesion in Narrative Speech, LSA coherence in Free Speech and LSA coherence in Narrative Speech (see Table 1 for results). The only one that reached significance was a positive relationship between LSA coherence in Free and Narrative Speech.

Table 1

Correlations between the four text measures

	Lex_Free	**Lex_Narr**	**LSA_Free**	**LSA_Narr**
Lex_Free	-			
Lex_Narr	.24 [*.202*]	-		
LSA_Free	.02 [*.906*]	.14 [*.463*]	-	
LSA_Narr	.03 [*.867*]	.30 [*.109*]	.41 * [< *.05*]	-

Note: LSA_Free, LSA coherence in the free speech task; Lex_Narrative, lexical cohesion in the narrative speech task; LSA_ Narrative, LSA coherence in the narrative speech task; Lex_Free, lexical cohesion in the free speech task. Table cells: r(28) [*p*]

* Significant at the .05 level

Finally, three main multiple regression analyses were run on FTD + and FTD − groups as subjects, with PANSS, CGI-S and GAF scores as dependent variables and, in all three of them, with lexical cohesion in Free Speech, lexical cohesion in Narrative Speech, LSA coherence in Free Speech and LSA coherence in Narrative Speech as predictors.

For the total PANSS score, both lexical cohesion in Narrative Speech (β = .429, $t(15)$ = 1.869, p = .081) and LSA coherence in Free Speech (β = .494, $t(15)$ = 2.068, p = .056) tended towards being positive predictors. When WAIS and TAP IQ were introduced in the model as predictors, in order to control for general cognitive ability, LSA coherence in Free Speech reached significance ($p < .05$) as a positive predictor, and lexical coherence in Narrative speech still showed a tendency towards significance (p = .052). The amount of variance explained in PANSS scores by the model (without IQ scales as predictors) was 16.9 %.

Further regression analysis using the same predictors were run with PANSS subscales as dependent variables. No significant predictors emerged for PANSS-P and PANSS-N subscales. However, they did for the PANSS-GP subscale, results can be seen in Table 2. Both lexical cohesion in Narrative Speech and LSA coherence in Free Speech were significant positive predictors, while LSA coherence in Narrative Speech showed a tendency to be a negative predictor.

Table 2

Results of the regression with PANSS-GP as the dependent variable

Predictor	**β**	***t*(15)**	***p***
LSA_Free	.573	2.615*	< .05
Lex_Narrative	.478	2.270*	< .05
LSA_Narrative	-.423	-1.831	.087
Lex_Free	-.347	-1.701	.110

Note: Predictors in order of contribution from top to bottom. LSA_Free, LSA coherence in the free speech task; Lex_Narrative, lexical cohesion in the narrative speech task; LSA_ Narrative, LSA coherence in the narrative speech task; Lex_Free, lexical cohesion in the free speech task.

* Significant at the .05 level

When IQ scores were introduced in the PANSS-GP model, both lexical cohesion in Narrative Speech ($p < .05$) and LSA coherence in Free Speech ($p < .05$) retained their status as positive

predictors, but the tendency for LSA coherence in Narrative Speech was attenuated (p = .107). The total variance in PANSS-GP scores (without IQ) was 30.1 %.

No significant predictors emerged in the models with CGI-S and GAF scales as dependent variables.

Discussion

Group differences

First, we will discuss the lexical cohesion results. According to the results of the ANOVA, lexical cohesion varied between groups. Such variation seemed to be primarily guided by healthy controls making more use of lexical cohesion than patients with FTD, with no apparent differences between tasks. One pairwise comparison, although not significant after correction, suggests that the tendency is for controls to also produce more lexical cohesion than patients without FTD. There was also no evidence of differences between the two patient groups. Therefore, there is no evidence to support hypothesis 1, which was that the FTD group would score higher than the other two groups. In fact, the results point in the opposite direction, with the FTD group producing less lexical cohesion than controls.

One possible reason for this failure to find the expected results might be the tasks in this study. Rochester & Martin (1979)'s free speech task was an interview in which "the subject was asked to choose the topics from anything he or she found interesting." (p.62). This may contrast with our Free Speech task, in which the topic was fixed and, since it requires talking about events in one's life, may elicit a more narrative speech. In fact, in the original findings by Rochester & Martin (1979) speech context had a big effect on lexical cohesion, explaining 20 % of the variance. The expected differences were only seen in their free interviews, while, in the narrative context, controls actually showed a higher amount (although not significantly) of lexical cohesion than the other two groups, like in our results. All this, together with the fact that there was no significant task effect in our data, points to the fact that our "Free Speech" task was more akin to a narrative one, similar to the other one. A true free speech task may provide a more unconstrained context, in which automatic associations between words are more readily observed, whereas the influence of this associative activity may not be detectable in more constrained contexts, like in this study.

With respect to the LSA results, the amount of coherence didn't seem to differ across groups or across tasks, and there was no interaction. It is true that pairwise comparisons show that controls are more coherent in narrative than in free speech, and, somewhat in line with our hypothesis 2., controls displayed higher coherence than patients without FTD during narrative speech. However, the fact that these differences disappear when corrections are applied and that they are observed in the absence of any main or interaction effects strongly suggests that they are due to chance. Given that groups didn't significantly differ, and there was no evidence that the FTD group had lower coherence scores, hypothesis 2. receives no support either.

Many different methodological reasons may account for this failure to find the expected results. LSA was used instead of more modern word embedding systems, like GloVe or word2vec, which, according to Iter et al. (2018), have shown to outperform LSA in some tasks. A second reason is the absence of any vector weighing scheme to obtain the sentence and text embeddings in our methodology. These schemes are mathematical procedures to obtain vectors for groups of words out of the vectors for the individual words, more elaborate than simply summing or averaging the vectors. For example, the TF-IDF weighting consists in multiplying each vector by the number of times the word appears in the current sentence and dividing by the number of documents it appears on in the original corpus, prior to averaging all vectors. Both Iter et al. (2018) and Just et al. (2019) fail to find any differences between the groups without using these schemes. As they point out, not using them makes word repetitions and longer sentences unduly increase coherence. Not applying this correction could be behind the negative results of this study. One final difference is the procedure used to generate the coherence scores. Previous studies had obtained cosines between each sentence and the original prompt by the interviewer (Elvevåg et al., 2007) or between each adjacent sentence produced by the participant (Bedi et al., 2015). In this study, we used sentence to full text comparisons. Although this method has been shown to produce differences between texts purposefully designed to be more or less coherent (McNamara et al., 2010) perhaps it is not sensitive enough to capture more subtle differences in coherence between our groups.

Relationship between cohesion and coherence

Correlations between all lexical cohesion and LSA coherence variables were performed as an exploratory analysis, without any specific hypothesis. One finding is that, while LSA

coherence scores are correlated across tasks, lexical cohesion scores are not. This may mean that whereas the use of lexical cohesion is very task dependent (as reported by Rochester & Martin, 1979), LSA coherence scores can be expected to be more stable within a participant, irrespective of the task they face.

Additionally, no evidence of a relationship was found between lexical cohesion and LSA coherence scores. This might be somewhat surprising because, as commented in the introduction, both measures rely on semantic connections between parts of the text. The absence of any relationship might provide a reason to believe in the independence of the two. However, a relationship might be found with a bigger sample, and we can only conclude that this data offers no evidence for it.

Prediction of symptomatology, severity and functioning

First of all, none of the coherence or cohesion scores were able to predict the severity of the disorder or adaptive functioning. Although this was predicted in hypotheses 3 and 4, no previous evidence existed for lexical cohesion, and the evidence for LSA coherence predicting functioning (Holshausen et al., 2014) was based on the analysis of verbal fluency responses and not coherence scores. For all we know, it might well be that neither of these measures are sensitive enough to predict severity or functioning.

Some predictors did emerge for symptomatology measured with the PANSS scale, and this result seemed to be guided by the prediction of the PANSS-GP scale. It is not clear why our measures would explain variance in the general psychopathology scale, which includes the symptoms less specific to psychosis. If anything, one might expect measures related to FTD to be predictive of positive symptomatology, but no such result was found. Perhaps coherence and cohesion disturbances found in our data are the product of deficits in attention, anxiety or other unspecific symptoms affecting performance, instead of being related to FTD per se.

With regards to the specific predictors, there was, first, a tendency for LSA coherence in the narrative section to be a negative predictor of general psychopathology. This would be in line with hypothesis 4 However, it wasn't close to significance in the prediction of the general PANSS scale, it was only a tendency in PANSS-GP, and since it lost its status as a clear tendency when IQ scores were controlled for, this result was probably a product of chance or an artifact of general cognitive ability. Two other predictors emerge more clearly. First, LSA coherence during free speech was a *positive* predictor of general psychopathology. The fact that higher coherence scores predicted higher amounts of symptomatology, the opposite of

what was expected, leads us to reject hypothesis 4 and is very hard to interpret. Second, lexical cohesion during narrative speech also positively predicted general psychopathology. This is what was expected, but it is difficult to reconcile with the findings of group differences, since lexical cohesion was found to be decreased for FTD patients compared to controls. Hypothesis 3., then, receives some limited support. Perhaps, during more narrative speech (see the discussion above on the tasks in this study) healthy controls show more lexical cohesion as a "normally used" tie, but when levels of cohesion are abnormally low (in the groups with patients) differences do have to do with pathological repetition, or perseveration.

One final reason may account for these last findings. FTD is not a unidimensional construct and may be divided in Positive FTD (comprising the more fluent, disorganized aspects) and Negative FTD (comprising poverty of speech). These two aspects have different neuropsychological correlates, with inhibition being exclusively related to Positive FTD (Bora et al., 2019). Qualitative inspection of the interviews with FTD patients in our sample suggests that many of them were more related to Negative FTD. Perhaps the expected findings, with LSA coherence negatively predicting symptoms, would be found if only patients showing evidence of Positive FTD were included in this study. The problems with inhibition in this section of FTD patients may be what gives rise to tangentiality, derailment, etc., FTD signs that may be what is captured by LSA incoherence measures.

Overall, however, the contradictory nature of many of our findings does not make them amenable for a clear theoretical explanation. They are more likely explained by methodological reasons.

Conclusions

In this study we explored the idea that lexical cohesion and LSA coherence, despite both depending on semantic associations within the text, may be independent and dissociable in the same sample, and each explain a different portion of the variance in symptomatology, severity of illness and functioning of patients. The results didn't support our hypotheses. Group differences weren't found for LSA coherence, and for lexical cohesion they went in the opposite direction to the one expected (with FTD patients showing decreased levels). Severity of illness and functioning couldn't be predicted. While LSA coherence and lexical cohesion both explained independent variance of general psychopathology, higher levels of coherence were unexpectedly related to increased symptomatology, and the expected positive

relationship between lexical cohesion and symptoms is hard to interpret in light of the unexpected group differences.

Many limitations of our study may explain the negative results. First, as discussed, there were many methodological differences with previous literature. Our Free Speech task was more narrative in character than a truly unstructured interview in which participants decide the topic. Our LSA coherence measure had not been previously used in the FTD literature, and no weighing schemes were used to correct for repetitions and length of sentences. Second, the effect of some potentially relevant variables, like the extent to which our FTD sample consisted of patients with Positive or Negative FTD, went unexplored and uncorrected. Finally, a low sample size (n =10) was used. This may have resulted in low statistical power and caused false negative errors.

In the future, this study could be replicated while correcting for the mentioned methodological deficiencies. Additionally, it would be useful to explore other variables potentially related to our textual measures, such as semantic and executive neuropsychological tests. As Hitczenko et al. (2021) argue, maybe too much effort in the Natural Language Processing literature is devoted to attempting to classify participants by using a wide array of measures. Perhaps in the future a greater care should be taken to validate individual measures, such as specific models of LSA coherence. It would be useful to know the specific relationship of these measures to FTD dimensions, symptomatology, and cognitive variables. The specific circumstances when differences between groups are or, as in this case, aren't found should be clarified.

References

American Psychiatric Association. (2010). Diagnostic and statistical manual of mental disorders, text revision (DSM-IV-TR®).

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi org.ezproxy.frederick.edu/10.1176/appi.books.9780890425596

Andreasen, N. C. (1986). Scale for the assessment of thought, language, and communication (TLC). *Schizophrenia bulletin*, *12*(3), 473.

Barrera, A., Mckenna, P. J., & Berrios, G. E. (2005). Formal thought disorder in schizophrenia: an executive or a semantic deficit? *Psychological Medicine, 35*(1), 121-132.

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D. C., Copelli, M., & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *Npj Schizophrenia, 1*(1), 1-7.

Bora, E., Yalincetin, B., Akdede, B. B., & Alptekin, K. (2019). Neurocognitive and linguistic correlates of positive and negative formal thought disorder: A meta-analysis. *Schizophrenia Research, 209*, 2-11.

Chaika, E., & Lambe, R. A. (1989). Cohesion in schizophrenic narratives, revisited. *Journal of Communication Disorders, 22*(6), 407-421.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

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

Covington, M. A., He, C., Brown, C., Naçi, L., McClain, J. T., Fjordbak, B. S., Semple, J., & Brown, J. (2005). Schizophrenia and the structure of language: the linguist's view. *Schizophrenia Research, 77*(1), 85-98.

Crider, A. (1997). Perseveration in Schizophrenia. *Schizophrenia Bulletin, 23*(1), 63-74. https://doi.org/10.1093/schbul/23.1.63

Docherty, N., Schnur, M., & Harvey, P. D. (1988). Reference performance and positive and negative thought disorder: A follow-up study of manics and schizophrenics. *Journal of Abnormal Psychology, 97*(4), 437–442. https://doi.org/10.1037/0021-843X.97.4.437

Dwyer, K., David, A., McCarthy, R., McKenna, P., & Peters, E. (2014). Higher-order semantic processing in formal thought disorder in schizophrenia. *Psychiatry Research, 216*(2), 168-176.

Elvevåg, B., Foltz, P. W., Rosenstein, M., & DeLisi, L. E. (2010). An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of neurolinguistics*, *23*(3), 270-284.

Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia Research, 93*(1-3), 304-316.

Foltz, P. W. (2007). Discourse coherence and LSA. *Handbook of Latent Semantic Analysis, 167*, 184.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*(2-3), 285-307.

Gomar, J. J., Ortiz-Gil, J., McKenna, P. J., Salvador, R., Sans-Sansa, B., Sarró, S., Guerrero, A. & Pomarol-Clotet, E. (2011). Validation of the Word Accentuation Test (TAP) as a means of estimating premorbid IQ in Spanish speakers. *Schizophrenia research*, *128*(1-3), 175-176.

Guy, W. (1976). *ECDEU assessment manual for psychopharmacology*. US Department of Health, Education, and Welfare, Public Health Service ….

Halliday M.A.K., Hasan R. (1976) *Cohesion in English*. London: Longman.

Harvey, P. D. (1983). Speech competence in manic and schizophrenic psychoses: the association between clinically rated thought disorder and cohesion and reference performance. *Journal of Abnormal Psychology, 92*(3), 368.

Hitczenko, K., Mittal, V. A., & Goldrick, M. (2021). Understanding language abnormalities and associated clinical markers in psychosis: the promise of computational methods. *Schizophrenia Bulletin, 47*(2), 344-362.

Holshausen, K., Harvey, P. D., Elvevåg, B., Foltz, P. W., & Bowie, C. R. (2014). Latent semantic variables are associated with formal thought disorder and adaptive behavior in older inpatients with schizophrenia. *Cortex, 55*, 88-96.

Iter, D., Yoon, J., & Jurafsky, D. (2018). Automatic detection of incoherent speech for diagnosing schizophrenia. Paper presented at the *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic,* 136-146.

Jorge-Botana, G., Olmos, R., & Barroso, A. (2013). Gallito 2.0: A natural language processing tool to support research on discourse. Paper presented at the *Proceedings of the 13th Annual Meeting of the Society for Text and Discourse,*

Just, S. A., Haegert, E., Kořánová, N., Bröcker, A., Nenchev, I., Funcke, J., Heinz, A., Bermpohl, F., Stede, M., & Montag, C. (2020). Modeling incoherent discourse in non-affective psychosis. *Frontiers in Psychiatry, 11*, 846.

Just, S., Haegert, E., Kořánová, N., Bröcker, A., Nenchev, I., Funcke, J., Montag, C., & Stede, M. (2019). Coherence models in schizophrenia. Paper presented at the *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology,* 126-136.

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin, 13*(2), 261-276.

Kramov, A. (2020). Evaluating text coherence based on the graph of the consistency of phrases to identify symptoms of schizophrenia. *arXiv Preprint arXiv:2005.03008,*

Kuperberg, G. R., Weber, K., Delaney-Busch, N., Ustine, C., Stillerman, B., Hämäläinen, M., & Lau, E. (2019). Multimodal neuroimaging evidence for looser lexico-semantic networks in schizophrenia:Evidence from masked indirect semantic priming. *Neuropsychologia, 124*, 337-349. https://doi.org/10.1016/j.neuropsychologia.2018.10.024

Landauer, T. K., & Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review, 104*(2), 211-240. https://doi.org/10.1037/0033-295X.104.2.211

Manschreck, T. C., Maher, B. A., & Ader, D. N. (1981). Formal thought disorder, the type-token ratio, and disturbed voluntary motor movement in schizophrenia. *The British Journal of Psychiatry*, *139*(1), 7-15.

McKenna, P. J., & Oh, T. M. (2005). *Schizophrenic speech: Making sense of bathroots and ponds that fall in doorways*. Cambridge University Press.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*(4), 292-330.

Minzenberg, M. J., Ober, B. A., & Vinogradov, S. (2002). Semantic priming in schizophrenia: a review and synthesis. *Journal of the International Neuropsychological Society*, *8*(5), 699-720.

Panicheva, P., & Litvinova, T. (2019, November). Semantic Coherence in Schizophrenia in Russian Written Texts. In *2019 25th Conference of Open Innovations Association (FRUCT)* (pp. 241-249). IEEE.

Rabagliati, H., Delaney-Busch, N., Snedeker, J., & Kuperberg, G. (2019). Spared bottom-up but impaired top-down interactive effects during naturalistic language processing in schizophrenia: evidence from the visual-world paradigm. *Psychological medicine*, *49*(8), 1335-1345.

Ragin, A. B., & Oltmanns, T. F. (1986). Lexical cohesion and formal thought disorder during and after psychotic episodes. *Journal of Abnormal Psychology, 95*(2), 181.

Roche, E., Creed, L., MacMahon, D., Brennan, D., & Clarke, M. (2015). The epidemiology and associated phenomenology of formal thought disorder: a systematic review. *Schizophrenia bulletin*, *41*(4), 951-962.

Rochester, S. R. & Martin, J. R. (1979). Crazy talk: A study of the discourse of schizophrenic speakers.

Rochester, S. R., Martin, J. R., & Thurston, S. (1977). Thought-process disorder in schizophrenia: The listener's task. *Brain and Language*, *4*(1), 95-114.

Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., & Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, *304*, 114135.

Voleti, R., Liss, J. M., & Berisha, V. (2019). A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE Journal of Selected Topics in Signal Processing, 14*(2), 282-298.

Wykes, T., & Leff, J. (1982). Disordered speech: differences between manics and schizophrenics. *Brain and Language, 15*(1), 117-124.

Zhang, M., Geng, L., Yang, Y., & Ding, H. (2021). Cohesion in the discourse of people with post-stroke aphasia. *Clinical Linguistics & Phonetics, 35*(1), 2-18. https://doi.org/10.1080/02699206.2020.1734864

Annexes

Annex I. Sentence division criteria

Criteris de separació en frases

- No són frases els nominals penjats. Per exemple:

i. T: *Los redactores clandestinos.*

P: Si, ocultos (…)

ii. *solo* (…) *los sitios peligrosos*

iii. P: Pues eso no lo sé, (51) pero (…) *personas y de esto, conocidos y familia* (…), *familia mía* [corpus Moya]

- No són frases tampoc les seqüències amb verb que no arriben a constituir una oració. Per exemple:

Y tienes que (…), *pero* (…) *como una enfermedad eso,*

Parataxi i coordinació

La parataxi amb repetició del tipus il·lustrat immediatament a sota serà considerada com a se- qüència de frases independents:

Quisiera estar en aliciente de vida (n), *quisiera tener un aliciente* (n+1), *quisiera encontrar en las cosas un aliciente* (n+2) [corpus Moya]

- Dues oracions són **frases independentes** si tenen

- subjectes gramaticals diferents amb un tòpic idèntic:

Y fueron$_i$ (i) *y les envenaron*$_j$ (ii) [exemple corpus fidmag]

- subjectes gramaticals idèntics amb un canvi de tòpic:

Entró por la ventana (i) *y al lobo no lo vio* (ii) [exemple inventat] *Yo ahora tengo esto* (46), *el equilibrio lo tengo* (47) [corpus Moya] [Amb parataxi en lloc de coordinació]

- Dues oracions son **una sola frase** en aquests dos casos:

- cas canònic: subjectes correferents

Entró por la ventana y se acercó a la puerta

- *gapping* en el segon conjunt coordinat (o paratàctic) tot i que els subjectes siguin diferents

La bolsa no acompaña y los días tampoco porque tengo unos horarios muy malos

Narració amb discurs directe

El discurs directe introduït pel verb *dir* o similar s'ha d'analitzar independentment del verb en qüestió, i l'oració en què apareix no podrà considerar-se deficient, agramatical.

Y entonces le dijo (i) *¿Dónde vas Caperucita?* (ii)

Altra casuística

- **(…)** Una frase no podrà incloure cap marca d'aquestes, i.e. una pausa molt llarga

- || **[…]** || No s'ha de comptar com a frase. Entrarà si de cas a l'anàlisi de la fase 3.

- Fórmules fossilitzades com *Vamos a ver; pues, no sé* no les comptarem com a frases tret que siguin marcadament anòmales.

- Les frases de l'entrevistador s'han de numerar però no analitzar.

Annex II. Annotation criteria

Criteris anotació cohesió FTD (Moya study 2)

ÍNDEX

Aquest document és un manual d'anotació per a l'estudi Moya 2, que essencialment presenta i desenvolupa els criteris d'anotació per a la primera fase de l'estudi que segueix al peu de la lletra (i expandeix en algun punt) la proposta i anàlisi per **lligats cohesius** (**cohesive ties**) de Halliday & Hasan (1976) aplicada per primer cop al Formal Thought Disorder per Rochester & Martin (1979).

Això no obstant, en l'apèndix 1 es presenten els criteris d'anotació per a les fases 2 i 3.

Què són els lligats cohesius?

Els lligats cohesius (LC) són aquells que s'estableixen entre un ítem que en pressuposa un altre pertanyent a una frase diferent del discurs, generalment prèvia —també és possible però molt més infrequent la catàfora, en què l'ítem pressuposat apareix posteriorment. Els LCs de H&H són tots endofòrics i entre frases. Serien de fet els responsables que el component verbal d'un discurs se sostingui quan ja no són d'aplicació les regles i principis que organitzen les frases internament.

Ni l'exòfora ni els ítems no fòrics poden entrar en LCs. L'exòfora no hi entra perquè remet (pressuposant-lo) al context situacional i no pas al verbal. Així, en una anàlisi de lligats cohesius a la H&H, cal deixar fora la dixi (de persona (*jo*, *tu*, etc.), lloc (*aquí*, *allà*, etc.) i temps (*ara*, *avui*, *demà*, etc.)), que és per defecte referència exofòrica. Quan la dixi sigui endofòrica, però, s'haurà d'anotar —serà previsiblement molt marginal que així sigui. Els ítems no fòrics no hi entren tampoc pel fet que simplement no pressuposen: no pressuposen ni en el context verbal ni en el context situacional. En absència de relació lèxica amb un ítem anterior (reiteració tal i com la defineixen H&H, o pertanyença a una mateixa família lèxica (col·locació per H&H), són exemple d'ítems no fòrics els sintagmes nominals (DP/NP) no definits (quantificats amb l'article indefinit o altres quantificadors no comparatius) i els genèrics, que en espanyol i les llengües romàniques sovint es presenten formalment com a definits, especialment quan són preverbals (cf. L'aigua *és bona per als ronyons* vs.

Beure aigua *és bo per als ronyons*). En connexió amb com (no) s'anota la genericitat, vg. la subsecció sobre els articles a la secció 3 —així com l'apèndix 1 per a la fase 3.

Aclarits els fonaments de l'anàlisi de H&H, tot seguit, a la secció 2, es descriuen, i s'il·lustren si cal, els diferents tipus d'LCs amb una adaptació a l'espanyol per al tipus R(eferència) i E(l·lipsi). Seguidament, a la secció 3, es presenten unes generalitzacions rellevants per a l'establiment dels LCs formulades sobre el comportament al respecte dels diferents tipus de categoria gramatical en espanyol. A continuació s'hi recullen unes observacions miscel·lànies d'interès. I, finalment, en l'apèndix 2 —després de presentar l'anotació per a les fases 2 i 3 en l'apèndix 1—, surten aplegats i exhaustivament anotats tots els exemples anotats prèviament de forma parcial, precisament per ressaltar el tipus de cohesió en discussió.

Lligats cohesius: criteris (fase 1 de l'estudi)

Els LCs es presenten en 5 tipus majors diferents i s'estableixen entre l'ítem cohesiu (a la columna 3 del full d'anotació) i un altre aparegut en una frase generalment

precedent (columna 7). La distància, anotada a la columna 6, pot ser 0 o altrament. És 0 quan l'element pressuposat apareix en la frase immediatament anterior. Si no és 0, la distància serà no mediada (Nx) i/o mediada (My). Així N3 voldria dir que hi ha 3 frases *entre* els dos extrems del lligat en què no apareix cap element mediador; i M2, que n'hi ha 2 que sí que contenen mediadors. N i M són independents: hi ha lligats a distància N, lligats a distància M i lligats a distància N+M. Els casos de catàfora, i.e. amb l'ítem cohesiu precedint el pressuposat, s'anoten com a K. Vg. H&H 1976: 339.

Els 5 tipus majors són el següents: **cohesió Lexica (L), Referència (R), Conjunció (C), Substitució (S) i El·lipsi (E)**. A continuació es presenten els corresponents subtipus així com l'aplicació del criteri de distància en els casos en què calgui. Per a R i E, com ja s'ha dit, els codis s'han adaptat a l'espanyol.

COHESIÓ LÈXICA: L

Són lligats cohesius L aquells en què l'element cohesiu i el pressuposat es relacionen pel contingut lèxic. Són 5, els tipus d'L. D'L1 a L4 es tracta de **reiteracions** (repeticions i ocurrències estretament relacionades des d'un punt de vista lèxic). L5, en canvi, recull les **col·locacions**, casos on la relació lèxica és patent però no tan íntima i sistemàtica com en les reiteracions. Els elements funcionals no entren mai en relacions L.

TIPUS

- Repetició del contingut lèxic L1
- Sinònim o similar amb hipònims inclosos L2
- Hiperonímia L3
- Ítem molt general L4
- Col·locació o relació lèxica diferent d'L1-L4 L5

L1: Repetició del contingut lèxic

L1 no comporta L16: vg. més avall, en aquesta mateixa secció L, com s'anoten els subtipus 6-9 d'L.

L2: sinònim o similar amb hipònims inclosos; **L3**, Hiperonímia

Respecte de L2 i L3, l'ítem cohesiu i no el pressuposat és qui determina si som davant d'L2 (en el cas d'hiponímia) o L3. Amb un exemple, si l'ítem cohesiu és *soroll* i el pressuposat és *so,* el tipus de cohesió és L2, mentre que si l'ordre d'aparició en el text fos al contrari i, doncs, *so* fos l'ítem cohesiu i *soroll,* el pressuposat, la relació seria L3. Exemples mig manllevats de H&H frases 14 i 16 p. 342-34.

L3: hiperonímia; **L4**: ítem molt general

Entre L3 i L4, la frontera pot ser molt difusa: per exemple, *sitio* respecte de *pendiente*, *arroyuelo*, etc. què és, L3 o L4? Aquí l'anotador/-a pot optar per la solució que cregui convenient i ser coherent amb l'opció triada.

L5

Per a L5, les col·locacions, no farem servir els subtipus 6-9 descrits a continuació. Seria massa complex, potser impracticable. Els casos dubtosos d'L5 se sotmetran a la decisió conjunta de l'equip anotador.

SUBTIPUS

Els subtipus 6-9 impliquen REFERÈNCIA (R) [≠ identitat o proximitat lèxica] i la referència no es llegeix mai directament en el contingut lèxic, sobre L, sinó a la perifèria funcional dels sintagmes nominals (SN/SD) referencials.

- Referència idèntica 6
- Referència inclusiva 7
- Referència exclusiva 8
- Referència no relacionada 9

Per exemple, suposem que tenim la frase *No solíamos ir a más **sitios*** en una conversa en què abans s'hagués dit *Íbamos al club*. L'ítem cohesiu *sitios* és L3 o L4 respecte de *club*, l'ítem pressuposat; però indubtablement el subtipus seria 8 ja que la *referència* de *no+más sitios* de l'ítem cohesiu és necessàriament exclusiva respecte de *el club.*. S'anotaria, doncs, com a L3.8 o L4.8.

Respecte de 6, cal tenir present que **L1.6** equival a pronom personal (vg. H&H: 283) o a un demostratiu locatiu: un pronom personal (R) podria aparèixer en lloc d'un L1.6 amb un nom propi gentilici; i un demostratiu com per exemple *allí* (també R) en lloc d'un L1.6 amb un nom propi locatiu. Vg. la diferència a (1) i (2), respectivament —vg. l'especificació d'R per a aquest exemple en particular a l'apartat final i, en general, a la secció dedicada a Referència (R). Noteu que els possessius confereixen al nominal repetit que els segueix una qualitat d'L1.6 també —fins i tot si són de primera o segona persona i, doncs, en general díctics i per això, com a díctics, no anotables.

4. – El otro día vino *Maria* con *tu primo*
  - Pero no estaba enferma *Maria*? (L1.6)
  - Pero no estaba enferma *ella*? (R)

- Pero no estaba en Londres *mi primo*? (L1.6)
- Pero no estaba en Londres *él*? (R)

5. – ¿No vivía usted en Barcelona en aquella época?
   - Sí, vivía *en Barcelona* (L1.6)
   - Sí, vivía *allí* (R)

L1.6 s'obté, doncs, en aquests casos:

- amb ítems cohesius que siguin noms propis gentilicis o locatius sempre que siguin referencials, tot i que en general, quan ho són, és preferible un pronom (H&H: 281)

No ho és *Tatiana* a *Se llamaba Tatiana* ni *Olga* a *Mi madre es la princesa Olga* encara que abans haguessin aparegut dins de SDs referencials. Són L1 i prou en no ser referencials sino SN/ SDs que fan de predicat.

Ex. Para cuando llegó Olga, la princesa, […] Mi madre es *la princesa* (L1) Olga (L1)

- amb SDs amb un determinant possessiu díctic (1a, 2a persona) —és l'estratègia de P03— sempre que es tracti d'SDs referencials, com passa únicament en el primer dels dos exemples següents —en el segon *tu madre* és predicat :

mi madre […] *mi madre* vino (L1.6) *vs*. tu madre …. ella *es tu madre* (L1)

Més sobre 6-9 per a L

Els casos d'L1-L4 en què no hi hagi un element funcional introductor que sigui de tipus R no rebran subespecificació 6-9. Això s'aplica naturalment als genèrics, als nominals amb funció de predicat i als nominals quantificats no comparatius. Així, les ocurrències de *(des)equilibrio* de P01 tant si apareixen sota la forma *un (des)equilibrio* com si apareixen sota la forma *el (des)equilibrio* no rebran rebran cap subespecificació. Els genèrics es comenten amb més detall en les observacions de l'apartat següent, el de Referència.

DISTÀNCIA

Per a un ítem cohesiu que sigui L1 respecte d'un altre, compten com a M les frases que continguin L2-L5 relacionats entremig. Altrament, per a L2-L5, només l'ocurrència precedent pressuposada es compta.

Atès que L no és referencial, no comptarem com a M, en canvi, cap frase amb un ítem R que pressuposi la L que s'analitza.

CARÀCTER NOMINAL O NO NOMINAL D'L, ±N —columna inexistent en el sistema de H&H

Quant a la columna afegida ±N, cal distingir els casos en què L consta d'un sol ítem i els casos en què la relació de cohesió L és d'algun tipus de repetició, i doncs L1, que inclogui més d'un ítem.

- Quan L afecta un sol ítem, considerarem +[N] tots els casos d'SD/SN amb N realitzat sigui l'SD/SN referencial o predicatiu. En casos d'L amb verbs lleugers (*hacía ruido, tenía sueño,* etc.) comptarem L com a +[N] però no pas en casos d'L en què el verb és més carregat. Anotarem com a –[N] tots aquells casos d'L en què L no és nominal.
- Quan la repetició, i, doncs, L1, afecta més d'un ítem, caldrà distingir al seu torn aquests altres dos casos:

➥ **Pura repetició**, amb els casos i l'anotació que s'especifica a continuació:

- repetició d'un SX complex que contingui elements nominals i d'altre tipus: +– si SX

= SN; –+ si SX ≠ SN

- repetició de dos o més elements no nominals: – –
- repetició de dos o més elements nominals: ++

Exemples:

6. Porque los huesos, las articulaciones (, los tendones) son lo más importante […] […] los huesos, las articulaciones (, los tendones) me duelen mucho: (L1, ++)

7. me duelen los huesos, las articulaciones (, los tendones) […]

[…] me duelen los huesos, las articulaciones (, los tendones): (L1, – +)

8. vienen, miran (y se van) […]

[…] vienen, miran (y se van): (L1, – –)

➥ **Repetició tipus plantilla**, però parcial lèxicament. Farem servir aquí la mateixa tipologia just esmentada, i.e. –+, +–, – –, ++. Ara, en aquest cas, l'element nou s'anota ulteriorment si així ho requereix —**es tracta que un mateix element no entri en una relació L dues vegades en la mateixa frase**. Compareu els tres exemples anteriors amb (6), (7) i (8):

9. Porque los huesos, las articulaciones, los tendones son lo peor […] […] los huesos, los músculos, los tendones me duelen mucho: (L1, ++)

(L5, +) per a *músculos* en relació amb l'estructura pressuposada inicial a (6)

10. me duelen los huesos, las articulaciones, los tendones […]

[…] me machacan los huesos, las articulaciones, los tendones: (L1, – +)

(L5, –) per a *machacan* en relació amb l'estructura pressuposada inicial a (7)

[…] me duelen los huesos, los ligamentos y los músculos: (L1, – +)

(L5, +) per a *ligamentos*/*músculos* en relació amb l'estructura pressuposada de partida (7)

11. vienen, miran, y se van […]

[…] se acercan, miran, y se van: (L1, – –)

(L5, –) per a *se acercan* en relació amb l'estructura pressuposada a (8) […] se acercan, miran, y se marchan: (L1, – –)

(L5, –) per a *se acercan/se marchan* en relació amb l'estructura pressuposada a (8).

Observació. Cal recordar que les repeticions (L1), les de plantilla incloses, són lèxiques. Si un pacient repeteix elements gramaticals (pronom *jo*, p. ex.; o el díctic *aquí*) no hi ha L de cap tipus. Per la mateixa raó quan la repetició és encapçalada per una preposició, no considerarem la preposició. Així

para fregar seria –N

en la cocina, +N

para fregar la habitación, –+N

sin ganas de hacer nada, +–N

En conseqüència, una repetició en què s'inclogui un **adjectiu** presentarà diverses possibilitats, totes de més d'un ítem:

harto de vivir aquí, *– – sitios retirados*, *+– orgulloso de su carrera*, *–+*

REFERÈNCIA: R

Són lligats cohesius R aquells en què l'ítem cohesiu i el pressuposat estableixen una relació de correferència en què la forma lingüística del sintagma pressuposat (o l'inicial de la cadena de correferència quan n'hi ha més d'un) i la de l'ítem cohesiu són necessàriament diferents, almenys pel que fa a la categoria funcional introductora d'ambdós. La correferència en sentit estricte demana definitud/especificitat. Ara, H&H inclouen també dins d'R aquells casos en què no hi ha correferència però, similarment al que passa amb la correferència, la categoria funcional introductora de l'ítem cohesiu només és interpretable endofòricament. Aquest segon cas correspon al tipus 4 en els tipus descrits a continuació.

TIPUS, adaptats a l'espanyol

Pronoms personals

singular, [+animat] R11

singular, [– animat] R12

plural, [+animat] R13

plural, [– animat] R14

singular, [+proposició] R15 Ex. *No* lo *sabían*

Demostratius

singular, [+animat] R21

singular, [– animat] R22

plural, [+animat] R23

plural, [– animat] R24

singular, [+proposició] R25 Ex. *Pero* eso *no ocurre nunca*

invariable lloc (*aquí, allí*) R26

invariable temps (*entonces*) R27

Articles

singular, [+animat] R31

singular, [– animat] R32

plural, [+animat] R33

plural, [– animat] R34

Comparació

identitat R41

similaritat R42

diferència R43

quantitat R44

qualitat (comp. i superl.) R45

Sobre 4, insistim-hi, no n'hi ha prou que hi hagi quantificació: la comparació endofòrica hi ha de ser. Considerem els exemples següents:

12. – Y tuvieron hijos
    - o *Muchos hijos* (L1)

A la segona proferència hi ha una el·lipsi [*tuvieron*] que es tractarà més endavant, i un L1 a *hijos*. *Muchos* no és fòric i no s'anota. Que no ho és ho mostra que [*Tuvieron*] *muchos hijos* funcionaria sense la proferència inicial. Consegüentment, no hi ha R i, doncs, cap subespecificació 6-9 del nom *hijos*, que en ser un nom comú no és referencial per si sol. Comparant (9) amb (10), el caràcter no fòric de *muchos* queda encara més clar. A (10), a part de l'el·lipsi associada amb *no,* diferent de la de l'exemple *supra* —vg. l'anàlisi més avall també—, ens trobem amb *más*, un quantificador que pel fet de ser comparatiu és R44.

13. – Y tuvieron tres hijos

– No, tuvieron *más* (R44) hijos (L1.8) [vg. H&H 1976: 342, anàlisi frase 14 [?]i explicació pàg 283]

DISTÀNCIA

Per als elements que estableixen lligats R, el procediment per establir el punt terminal consisteix a retrocedir fins que es trobi una ocurrència amb el descriptor lèxic corresponent, que no serà necessàriament l'inicial. Les frases entremig que continguin pronoms correferents amb el que s'analitza seran M; altrament, N.

Substitució: S

La substitució es dona quan l'ítem cohesiu reemplaça el pressuposat de forma que restituint el pressuposat en el lloc del cohesiu el resultat és equivalent en significat. És per tant una relació del substitut amb la forma lingüística pressuposada pròpiament

dita (amb el *wording*), cosa que no ocorre amb R. Amb R, la restitució de l'element pressuposat en el lloc de l'ítem cohesiu fa que desaparegui la (co)referència que es busca. H&H diuen que S és una relació entre ítems lingüístics (nivell lexicogramatical) mentre que la referència és una relació de significat (nivell semàntic). Els exemples

14. i (12) ens mostren la diferència entre substitució i referència, respectivament.

15. Substitució
- Póngame *unas patatas*
- A mí póngame *lo mismo* = A mí póngame *unas patatas*

16. Referència

_ Tengo *unas patatas* de sobras

- Dáme*las*. *Las* aprovecharé ≠ dame *unas patatas*

TIPUS

Nominal (S1)

[crec que no és un bon terme però respectarem l'original]

• de complement	*lo mismo*	**S11**	Ex. (13) i (14); també (11)
• d'atribut	*lo, lo mismo*	**S12**	Ex. (15)

17. – Defenderemos la justicia y la democracia por encima de las leyes establecidas

– Bueno, ellos defienden *lo mismo* (S11)

18. – Ayer sobraron muchas bebidas y un poco de pan

– Hoy sobró *lo mismo* (S11)

19. – Es maestra / Es genial

– Sí, *lo* es (S12)

Verbal (S2)

● de verb	*hacerlo*	**S21**	Ex. (16)
● de procés	*hacer lo mismo* (S*v*); *pasar lo mismo* (SV)	**S22**	Exs.(17)-(18)
● de proposició	*ser así*	**S23**	Exs. (19)

20. – El mes que viene me vacuno

– Yo *lo haré* en setiembre (S21)

21. – Compraremos algo de ropa en las rebajas
    - Yo *haré lo mismo* (S22)

22. – Ayer sobraron muchas bebidas y un poco de pan
    - Hoy *pasó lo mismo* (S22) cf. (14)

23. – Juan es un imbécil
    - *És así*/*Así es*. (S23) Llevas toda la razón

EL·LIPSI: E

L'el·lipsi és S(ubstitució) per Ø. Així, a diferència del que ocorre amb S, en una el·lipsi allò elidit ha de poder ser restituït sense més canvis (tret potser de fonològics —el cas de *uno*, examinat a sota), mentre que amb S cal treure el substitut mateix perquè la restitució sigui bona.

A tall d'il·lustració, abans d'especificar els subtipus d'el·lipsi rellevants per a l'espanyol, pot observar- se que el que en català s'expressa amb substitució nominal gràcies al clític *en*, en espanyol és el·lipsi (substitució per Ø). Ho il·lustren els exemples (20) i (21). Tant en espanyol com en català, una alternativa correcta a (20) i (21), seria l'el·lipsi del tipus il·lustrat a (22) —especificat més avall.

24. – Vols patates?
    - No, no *en* (S11) vull *més* (R44)
    - No, no (*en) vull més patates [≠ No, no en vull més, de patates]

[El clític *en* es comporta com *one* en anglès quan va precedit d'un element funcional quantificat i individuador però no definit, llevat, és clar, que *one* no és clític i apareix doncs en el mateix lloc que el sintagma que reemplaça]

25. – ¿Quieres patatas?
    - No, no quiero *más (R44) (E)*
    - No, no quiero más patatas

26. – Vols patates? / ¿quieres patatas?
    - Sí (E)
    - Jo també / yo también (E)

TIPUS

El·lipsi nominal (E1)

27. – Antes había muchos médicos ….
    - Ya, ahora, en cambio hay muy pocos (E1) Noteu que funcionaria igual invertint els SNs:

28. – Antes había muy pocos médicos
    - Ahora hay muchos(E1)

El cas de UNO. En espanyol, *uno* en un exemple com el següent el considerarem el·líptic encara que per a la restitució calgui treure -o.

29. – ¿Tiene muchos hijos, no? / ¿Tiene muchas hijas?
o Tiene uno / tiene una (E1)

Abans de considerar els altres tipus d'el·lipsi, la verbal i l'oracional, cal tenir present l'estructura de l'oració en anglès, fent atenció sobretot al grup verbal i a la separabilitat de l'auxiliar i verb en les formes analítiques o no sintètiques –*worked* és sintètica (o *simple* per *H&H)*; *had worked*, és analítica.

Part Modal	| Part proposicional		
Subjecte	**AUX | Verb Predicador/ grup verbal**	Complement	Adjunt(s)
The students	will | watch	a moovie	in the classroom

Com que en espanyol el grup verbal no és separable en dues unitats amb vida sintàctica pròpia, ni l'el·lipsi verbal (E2) ni l'el·lipsi oracional (E3) segueixen la mateixa pauta que en anglès on la separació en auxiliar i verb del grup verbal és, en canvi, determinant. Respecte de la implicació el·lipsi verbal → el·lipsi oracional és simplement el resultat del caràcter canònicament obligatori del grup verbal dins de l'oració. L'el·lipsi verbal, d'altra banda; es desglossa en a) i b). Vg. detalls a H&H (1976: 335)

a. el·lipsi operador (≈auxiliar) verbal → el·lipsi modal oracional
b. el·lipsi lèxica (≈verb) verbal → el·lipsi proposicional oracional

En definitiva, en espanyol no trobarem ni el·lipsi del verb amb presència de l'operador (o auxiliar) (E21) ni el·lipsi de l'operador (o auxiliar) amb presència del verb (E22) ni els corresponents implicats a nivell oracional (E31 i E32) derivats de la centralitat/obligatorietat del grup verbal a les oracions canòniques d'aquella llengua. Considerant que l'el·lipsi verbal i la implicada a nivell oracional a l'anglesa poden descartar-se de ple (E21, E22; E31; E32), aplicarem simplement E2 a casos de *gapping,* il·lustrat a (26), i E3 a casos d'el·lipsi penjada de la polaritat canònica explícita (*sí*, *no*) o implícita (*también*, *tampoco* i variants) que no puguin catalogar-se

d'E33 pel fet de no formar constituent amb un element precedent concorrent; i.e. en casos com els de (27). L'el·lipsi verbal i oracional queda per a l'espanyol així:

El·lipsi verbal (E2)

30. – Mario se compró un libro y yo, unos zapatos. ¿Y tu?
    - Yo un perfume (E2)

El·lipsi oracional (E3)

31. – ¿Vendrán los tres mañana? / ¿Han hecho los tres los deberes antes de acostarse?
    - Pedro sí (E3)
    - Juan también (E3) pero María no (E3)

A diferència dels casos E2 i E3 a seques comentats i il·lustrats aquí al damunt, els casos que segueixen es corresponen amb els de H&H (1976: 335) per a l'anglès. Es tracta de l'*el·lipsi oracional general*, E33, en què queda un únic constituent de l'oració; i l'el·lipsi oracional zero, E34, en què no en queda cap.

El·lipsi oracional general (E33)

• *qu*	**E33.1**	Ex. (28)
• sí/no	**E33.2**	Ex. (29)
• altrament, però 1 sol constituent	**E33.3**	Ex. (30)

Aquí els descriptors *qu, sí/no, altrament* fan referència a l'únic constituent que pivota l'el·lipsi.

32. – Vivía en Barcelona

– Con quién ? (E33.1.9)

33. – ¿Vivía en Barcelona entonces?
    - No (E33.2.6), había vivido allí antes

34. – ¿Dónde vivía usted?
    - En Barcelona (E33.3.7)

El·lipsi oracional zero (E34)

35. – Esta vez ganamos
    - ¿Estás segura ? (E34.9)

36. – Dice que ganaremos
    - No sé (E34.8), ya veremos (E34.9)

Noteu que tant en els exemples d'E33x com a E34 s'hi han aplicat les subespecificacions següents — que venen determinades per la frase prèvia que conté el material realitzat corresponent a l'el·lipsi.

- pregunta o resposta sí/no 6
- pregunta o resposta *qu* 7
- discurs indirecte 8
- altrament 9

El·lipsi no especificada (E)

Qualsevol altre tipus d'el·lipsi que no encaixi en els establerts aquí al damunt, s'anotarà simplement com a E.

CONJUNCIÓ: C

TIPUS

AdditiuC1

Adversatiu C2

Causal C3

Temporal C4

DISTANCIA

Generalment, però no sempre, és 0. Sigui com sigui, no l'anotarem

No es compten si no estableixen cap enllaç. En aquest cas els comptarem a la fase 2 com a erronis o fallits.

Categories gramaticals (funcionals i lèxiques). Criteris Articles

Casos d'articles formalment determinats que no entren en lligats R

- Els articles determinats que coocorrin amb **noms propis** gentilicis, locatius o similars **no entren** en lligats cohesius R(eferencials) perquè són expletius. Així, en una repetició de nom propi precedit d'article només comptarem un lligat L, concretament un L1.6 generalment —en una llista pot perdre la R. En canvi un SD definit que en pressuposi un altre definit o indefinit amb el mateix contingut nominal, comptem 2 lligats, un R3x i un L1y
- Els articles determinats que coocorren amb noms amb els quals contribueixen a lliurar una **interpretació genèrica no entren** en lligats cohesius R(eferencials). A diferència del cas anterior, la repetició lèxica no té força referencial per si sola i, doncs, anotarem L1 sense especificació ulterior. El mateix s'aplica als altres tipus major d'L que aniran sempre sense subespecificació si són genèrics.
    - Cal no confondre la genericitat amb l'expressió de la **possessió inalienable** que cal tractar com a díctica en el cas que el clític o la flexió que l'expressa així ho sigui (1a i 2a

persona gairebé sempre). La dixi coocorrent amb una repetició lèxica (L1) fa que aquesta passi a considerar-se del subtipus 6, tant amb possessius prenominals — vg. (33)

— com amb els clítics/flexió de la possessió inalienable —vg. (34):

37. – Mi madre me trajo aquí […]
    - Yo hablo con *mi madre* todos los días (L1.6)

38. – Me golpée la espalda […]
    - *Me* duele *la espalda* […] (L1.6)
    - Ya le he dicho, doctor, que teng*o* daño en *la espalda* (L1.6)

- Els articles determinats que apareixen acompanyant **noms amb funció no referencial, predicativa, no entren** tampoc en lligats cohesius i reben exactament el mateix tractament que els genèrics en el sentit que els nominals que entren en aquests sintagmes no s'han de subespecificar si són ítems cohesius de qualsevol tipus d'L.

Demostratius

Amb noms propis (*Ay este Guillermo que cara tiene!*) i genèrics (*Estos deberes!*), s'apliquen les observacions immediatament superiors.

Per als nostres codis R26 i R27, **díctics** de lloc i temps, cal tenir present que si són exofòrics no s'anoten en la fase 1 d'anàlisi dels lligats. Els **díctics exofòrics** (o endofòrics) manifestament incorrectes s'anotaran com a errors en la fase 2.

Quantificadors

L'article indefinit no entra en lligats cohesius com no hi entra tampoc cap quantificador que no sigui fòric. És la diferència entre *más* y *muchos* il·lustrada a (3) i (4) *supra.*

Pronoms

El pronom personal genèric **se (o uno)** no entra en lligats de cap tipus. No hi entra tampoc si la referència és a la primera persona —el cas de P15— ja que si bé aquí és referencial es tracta de referència situacional i no textual.

Noms

Els noms propis relacionats lèxicament que no corresponguin a participants amb un rol discursiu mínimament sostingut els comptarem com a casos d'L5. En canvi, els casos d'L1 amb noms propis seran L1.6 generalment —tot i que si apareixen amb article aquest no s'anotarà; vg. el primer punt de la subsecció *articles* en aquesta mateixa secció. Així, si un pacient introdueix una sèrie de noms de països, de gent famosa, sense reprendre'ls caldrà anotar aquests nominals com a L5 bé a la fase 1 (*between-sentence analysis*) bé a la fase 3 (*within sentence analysis*), allà on pertoqui.

Verbs

A banda del paper marginal que fan en la variable S, els verbs entraran principalment sota L. En els casos de repetició, i.e. L1, no hi pot haver subespecificació 6-9 amb un verb, ja que els verbs com a verbs (sense el temps) no fan referència.

Preposicions

No entren en el còmput L. Observeu l'anàlisi dels sintagmes següents en cas que fossin repeticions, i.-e. L1:

en una casa = *una casa*: L1 N +

en una casa encantada = *una casa encantada*: L1 N + –

para dormir = *dormir*: L1 N –

para dormir en invierno = *dormir en invierno*: L1 N – +

Observacions Qüestions tècniques

Les majúscules només per als codis L, R, S, E, C, etc.; qualsevol altre text, sempre en minúscula. Marqueu amb color les files corresponents a l'entrevistador

Distància:

S'anota per a L i R; la resta, no cal.

A més d'M i N, explicades més amunt, recordeu que hi ha catàfora (K).

La columna ±N s'anotarà tipogràficament [fixeu-vos en els espais]:

N +, p. ex. *mi tío*, *qué horror, cien pozos, jamón de york, otra hermana*

N –, p. ex. *me bañaba* (respecte de *me bañó*), *querido* (respecte de *me quería*)

N + –, per a casos d'SN complex amb un o més SX* inclosos, p. ex. *personas muy buenas, con el estropajo me pasó por aquí* —vg. el cas de les preposicions al final de la secció 3.

N – +, per a casos d'SX* complex amb un o més SN inclosos, p. ex. *bulle a mi alrededor una distracción y un aliciente*

N + +, per a casos d'SN complex amb un o més SN inclosos, p. ex. *una casa del congreso eucarístico*, *hotel Esplanade*, *duquesa Tatiana*,

N – – , per a casos d'SXcomplex amb un o més SX inclosos, p. ex. *porque no ne quería cuidar,*

SX* = Qualsevol sintagma no nominal (SD ≈ SN) [exemples del corpus Moya aquí]

Les tres anàlisis/fases per pacient han de quedar a la mateixa pàgina de l'excel, no en fulls diferents. Les separarem amb columnes en blanc.

Repeticions

Les repeticions intrafrasals que puguin considerar-se dèficits de fluència i, doncs, majoritàriament d'items funcionals no es computaran. Totes les repeticions han de tenir un ítem lèxic, amb contingut descritptiu.

Convencions del sistema d'anotació.

Cal reproduir-les talment al començament de cada entrevista anotada

T: Terapeuta (dr. Moya) P: Pacient

- … allargament de l'element vocàlic immediatament precedent
- (.), (..), (…): Silencis anòmals; com més llargs, més punts
- [.], [..], […]: Seqüències inintel·ligibles; com més llargues, més punts
- ||xyz||: Seqüències penjades, rellevants per a la fase 2.
- |xyz|: frases intercalades numerades, i.e. per anotar.

- Interrupcions, superposicions menors que no s'anoten, entre claudàtors directament en el text.
- Observacions d'interès entre claudàtors i mida de font 10, [xyz]
- Altres aclariments entre claudàtors i en format subíndex.

39. Apèndix 1. Sobre l'anotació de les fases 2 i 3 de l'estudi FASE 2, ANÀLISI D'ERRORS

Hi haurà 7 variables, dues d'elles subespecificades.

Failed Reference (FR)

FR1: Nominal (SD que conté un nom)

FR2: Pronominal, amb pronom manifest o no (pro)

FR3: Hanging NPs /SN penjats

- **Hanging XP (HP) /SX penjats** on X no pot ser (pro)nom
- **Failed Conjunction (FC):** connector oracional fallit

Clanging (CL)

- **Parafàsies (P)**

P1: fonèmica

P2: semàntica

Neologismes (N)

- **Gramàtica (GRAM)**

Errors gramaticals que no entren dins FR o HP

Quant a FR2, en casos com ara *Me dijo que lo sabia, Vinieron y me dijeron* comptarem un sol FR2 en el cas que el pronom flexionat principal (o inicial) sigui tal cosa, i.e., no sapiguem a qui fa referència. El mateix s'aplica entre frases. i.e. només anotarem un FR2 en una cadena que es resoldria sense error en el cas que la primera ocurrència fos referencialment inequívoca.

Els subjectes fonèticament buits de les oracions no inicials d'aquests dos exemples es compten com a R1 i R13 a la fase 3.

FASE 3, ANÀLISI DE LA COHESIÓ DINS DE LES FRASES

En aquesta fase s'anoten les mateixes variables que a la fase 1, dins però de les frases individuals. Això suposarà que, tret de la cohesió lèxica que pot apareixer en una frase simple (o mono- oracional), la resta de variables requereixin una frase complexa i, doncs, donada la poca freqüència d'aquest tipus de frase, que poques vegades apareixin.

Quant a la cohesió lèxica, L, mirem com s'afegiria a la identificada a la fase 1 en aquesta tercera fases sobre l'exemple (6), reintroduït aquí com (35):

40. – Porque los huesos, las *articulaciones*, los *tendones* son lo peor […] […] los huesos, los *músculos*, los *tendones* me duelen mucho.

Primera frase

- Fase 1: no s'anota res perquè és la primera vegada que apareixen aquestes peces lèxiques.
- Fase 3: L5, L5. Aquestes dues ocurrències d'L5 tenen com a ítems cohesius els elements en cursiva. Per a *articulaciones*, el pressuposat és necessàriament *huesos*, perquè és el primer en aparèixer; per a *tendones,* en aquest exemple (però potser no en d'altres) tant és que considerem el pressuposat *huesos* com *articulaciones*. La regla és que el terme pressuposat ha de precedir l'ítem cohesiu.

Segona frase

- Fase 1: L1++; L5 +, per a *músculos,* que és nou i, doncs, no inclòs en el còmput L1++, en relació amb l'estructura pressuposada inicial.
- Fase 3: Insistim-hi. L'anotació en aquesta fase es fa amb completa independència de les relacions establertes a la fase 1. Així a la segona frase de (35) hi ha dues ocurrències d'L5, *músculos* y *tendones*. La primera és entre *huesos* (pressuposat) i *músculos* (ítem cohesiu); la segona és entre *tendones* com a ítem cohesiu i *músculos* o *huesos* (indistintament aquí) com a element pressuposat.

Quant a R, es donarà generalment en una subordinada d'acord amb el principi B del lligam. S'entén, doncs, que els clítics del *clític doubling* obligatori de l'espanyol no s'anoten sota R. Així a *Le^{i} dije a María$_{i}^{i}$ que* pro$_{i}$ *se fuera* només la relació expressada en els subíndexos es compta com a R (R11). Per contra, una relació de correferència com la permesa a *Juan$_{i}$ vino con su$_{i}$ amigo,* seria un cas d'R en una frase simple —potser l'únic possible (i permès pel principi B). El cas segurament més freqüent serà del tipus il·lustrat a (36):

41. –El médico me dijo que se encargaría *él* (R11)

No són R els clítics de represa d'un dislocat. Així a la frase ¿*Y el conde de la Goteki l**o** ha oído nombrar*? no s'anotaria res a la fase 3.

Quant a C, forçosament caldrà una frase complexa articulada gràcies a C

42. – Lo sé *porque* él me lo dijo (C3)

Quant a S i E, és el mateix.

En definitiva, gairebé exclusivament L podrà 'cohesionar' una frase simple, mitjançant enumeracions o llistes, com a (35) o establint un lligat entre SD precopular i SD postcopulat d'una copulativa, etc., com a (38).

43. Mi madre son mis *monjas* y mis *maestras* (L5), (L5)

Cal remarcar que en la fase 1 anotaríem, en canvi, un sol ítem cohesiu L, si és que la frase (38) ja hagués establert una relació de cohesió lèxica amb una frase prèvia en què, suposem, hagués sortit *mi padre.* Anotaríem en aquest cas *mi madre* com a L5. En canvi, si a la frase prèvia l'ítem pressuposat fos *mis curas*, a la fase 1 anotaríem millor *mis monjas* com a L5. A la fase 3, però, aquestes relacions entre frases s'han d'ignorar del tot i, per tant, l'element cohesiu mai podrà ser l'inicial en un lligat cohesiu, tal i com ja s'ha explicat per a l'exemple (35) *supra.* A (38) hi ha, doncs, dues ocurrències d'L5. La primera és la del lligat entre *monjas* (ítem cohesiu) i *madre* (element pressuposat); la segona és entre *maestras* (ítem cohesiu) i *monjas* (element pressuposat). Noteu que en aquest cas, en contrast amb (35), *monjas* sembla pressuposar més clarament *maestras* que no pas *madre.*

Les columnes d'anotació seran idèntiques a les de la fase 1, excepte que òbviament no hi haurà columna per a la distància; en canvi, en tindrem una per als (pro)nominals que s'interpretin com a genèrics (GEN).

44. – El otro día vino *Maria* con *tu primo*

- Pero no estaba enferma *Maria*? (L1.6)
- Pero no estaba enferma *ella*? (R11)
- Pero no estaba en Londres *mi primo*? (L1.6)
- Pero no estaba en Londres *él*? (R11)

45. – ¿No vivía usted en Barcelona en aquella época?

- Sí, vivía *en Barcelona* (L1.6)
- Sí, vivía *allí* (R26)

46. Porque los huesos, las articulaciones (, los tendones) son lo más importante […]

L5+ (, L5+) fase 3

[…] los huesos, las articulaciones (, los tendones) me duelen mucho: L1, ++ fase 1 L5+ (, L5+) fase 3

47. me duelen los huesos, las articulaciones (, los tendones) […]

L5+ (, L5+) fase 3

[…] me duelen los huesos, las articulaciones (, los tendones): L1, ± fase 1 L5+ (, L5+) fase 3

48. vienen, miran (y se van) […]

L5– fase 3

[…] vienen, miran (y se van): L1, – – fase 1

(L5)– fase 3 [*L5 aquí és per* **se van***, d'aquí el parèntesi*]

49. Porque los huesos, las articulaciones, los tendones son lo peor […]

L5+, L5+ fase3

[…] los huesos, los músculos, los tendones me duelen mucho: L1, ++; L5, + per a *músculos* en relació amb (6), fase 1 [*noteu la diferència amb les repeticions completes de (3), (4) i (5)*]

L5+, L5+ fase3

50. me duelen los huesos, las articulaciones, los tendones […]

L5 +, L5+ fase3

[…] me machacan los huesos, las articulaciones, los tendones L1, ±; L5, – per a *machacan* —tot respecte de (7)— fase 1.

L5+, L5+ fase 3

[…] me duelen los huesos, los ligamentos y los músculos: L1, ±; L5, + per a *ligamentos*/*músculos* —tot en relació amb (7) fase 1.

L5+, L5+ fase 3

51. vienen, miran, y se van […]

L5, per a *se van* respecte de *vienen,* fase 3

[…] se acercan, miran, y se van: L1, – –; L5, – per a *se acercan* fase 1 L5 per a *se van* respecte de *se acercan,* fase 3

[…] se acercan, miran, y se marchan: L1, – –; L5, – per a *se acercan/se marchan* en relació amb la frase de partida (8)

L5 per a *se marchan* respecte de *se acercan*, fase 3

52. – Y tuvieron hijos
- *Muchos hijos* (L1) [Aquí no hi ha R; cf. (10)]

53. – Y tuvieron tres hijos
- No, tuvieron *más* hijos (R44) (L1.8)

54. – Póngame *unas patatas*
- A mí póngame *lo mismo* (S11)

55. – Tengo *unas patatas* de sobras
    - Dáme*las* (R14). *Las* aprovecharé (R14)

56. – Defenderemos la justicia y la democracia por encima de las leyes establecidas
    - Bueno, ellos defienden *lo mismo* (S11)

57. – Ayer sobraron muchas bebidas y un poco de pan
    - Hoy sobró *lo mismo* (S11)

58. – Es maestra / Es genial
    - Sí, *lo* es (S12)

59. – El mes que viene me vacuno
    - Yo *lo haré* en setiembre (S21)

60. – Compraremos algo de ropa en las rebajas
    - Yo *haré lo mismo* (S22)

61. – Ayer sobraron muchas bebidas y un poco de pan
    - Hoy *pasó lo mismo* (S22) cf. (14)

62. – Juan es un imbécil
    - *És així*/*Així es*. (S23). Llevas toda la razón

63. – Vols patates?
    - No, no *en* vull *més* (S11) (R44)
    - No, no (*en) vull més patates [≠ No, no en vull més, de patates]

[El clític *en* es comporta com *one* en anglès quan va precedit d'un element funcional quantificat i individuador però no definit, llevat, és clar, que *one* no és clític i apareix doncs en el mateix lloc que el sintagma que reemplaça]

64. – ¿Quieres patatas?
    - No, no quiero *más (R44) (E1)*
    - No, no quiero *más patatas* (R44) (L1.9)

65. – Vols patates? / ¿quieres patatas?

– Sí (E33.2.6)

- Jo també / yo también (E3)

66. – Antes había muchos médicos ….
    - Ya, ahora, en cambio hay muy pocos (E1)

67. – Antes había muy pocos médicos
    - Ahora hay muchos(E1)

68. – ¿Tiene muchos hijos, no? / ¿Tiene muchas hijas?
    - Tiene uno / tiene una (E1)

69. – Mario se compró un libro y yo, unos zapatos. ¿Y tu?
    - Yo un perfume (E2)

70. – ¿Vendrán los tres mañana? / ¿Han hecho los tres los deberes antes de acostarse?
    - Pedro sí (E3)
    - Juan también (E3) pero María no (E3)

71. – Vivía en Barcelona

– Con quién ? (E33.1.9)

72. – ¿Vivía en Barcelona entonces?

- No (E33.2.6), había vivido allí antes

73. – ¿Dónde vivía usted?
    - En Barcelona (E33.3.7)

74. Esta vez ganamos
    - ¿Estás segura ? (E34.9)

75. – Dice que ganaremos
    - No sé (E34.8), ya veremos (E34.8)

Totes dues considerades respecte de la frase no el·líptica de Dice que …, i.e. amb discurs

indirecte. Considerar la segona com a E34.9, com en l'ocurrència original en el text principal, suposa referir una el·lipsi, la de *ya veremos*, a una altra el·lipsi, la de *No sé.*

76. – Mi madre me trajo aquí […]
    - Yo hablo con *mi madre* todos los días (L1.6)

77. – Me golpée la espalda […]
    - *Me* duele *la espalda* […] (L1.6)
    - Ya le he dicho, doctor, que teng*o* daño en *la espalda* (L1.6)

78. Porque los huesos, las *articulaciones*, los *tendones* son lo peor […] (L5, L5) fase 3

[…] los huesos, los *músculos*, los *tendones* me duelen mucho: (L1, ++) fase 1; (L5, L5) fase 3 (L5, +), fase 1, per a *músculos* en relació amb l'estructura pressuposada inicial

79. –El médico me dijo que se encargaría *él* (R11)

– Lo sé *porque* él me lo dijo (C3)