INSTITUTO NACIONAL DE ASTROFÍSICA, ÓPTICA Y ELECTRÓNICA

TECHNICAL REPORT No. 696

COMPUTATIONAL SCIENCE DEPARTMENT

Computational Analysis of the Evolution of Atypical Language Caused by a Head Injury.

Marisol Roldán Palacios
Aurelio López López

March 21, 2025

Tonantzintla, Puebla.

Luis Enrique Erro No. 1 Sta. María Tonantzintla,
C.P. 72840, Puebla, México.

Computational Analysis of the Evolution of Atypical Language Caused by a Head Injury

Marisol Roldán Palacios Aurelio López López

Computational Science Department
National Institute of Astrophysics, Optics and Electronics
Luis Enrique Erro # 1, Santa María Tonantzintla, Puebla, 72840, México
E-mail: {marppalacios, allopez}@inaoep.mx

Abstract

Sequelae of language disorder as a result of head injuries is a problem that, beyond remaining stable, tends to increase, according to evidence reported in different studies. Therefore, language impairments require interdisciplinary attention.

From the perspective of Computational Science, aimed at providing information for a more precise characterization of these alterations, a solution involves different aspects. In particular, the limited data available due to the complexity in the collection of samples, the involved costs, or the insufficient availability of specialists.

When alterations in language competences are the subject of study, data extension or augmentation techniques are not applicable. So with scarce data, prevailing data mining techniques are not appropriate to examine it. Therefore, in the present investigation an approach with alternative methods to study the data is proposed.

We propose a trajectory representation of the narrative sample, after having identified the need to track the variability of the relationship among features over time. The aforementioned with the purpose of evaluating their contribution in the discrimination of language disorders analyzed, in addition to the aim of generating knowledge to comprehend language development.. The hypothesis that guides us is that, based on the study of the internal evolution of narrative instances, we can monitor this variability.

The analysis of features as a path also leads us to develop a comparison process in terms of proximity. This as part of the formulation of a complete technique for tracking this variation in the relationship among features, which comprise the recovery period primarily for instances of the narrative task of retelling a story.

Keywords: Feature relationship – Variability – Trajectory – Narrative –Language disorders – Head injury

Contents

1 Introduction

There is a global estimation of sixty-nine million cases of traumatic brain injury (TBI [1]) that can vary from 64 to 74 million, where numbers are annual approximations. An increasing global tendency in these figures places TBI as a leading cause of death, disabilities, or a condition of special needs [2, 3, 4]. Furthermore, that approximation does not include events coming from the phenomenon called *silent epidemic* where, for different circumstances, TBI cases are not accounted [5]. Depending on the complexity of the effects following a brain injury, the emotional and behavioral impact could represent a risk for acquiring psychological disorders [6]. Among moderate-severe TBI cases, deficits of awareness, reasoning, and language can persist in around 65% of patients [7]. These records show the need for further research to provide additional information from specific studies, specially in language impairments.

There is an accumulated number of works to understand ordinary language from heterogeneous approaches. Though, the line working on atypical language [2] is quite minor. Besides, the efforts to study the restructuring on language as a consequence of head trauma during the recovery stage are limited [8], and those incorporating computational resources are restricted to a few. In part, given that in general, data for analyses on impaired restructured language is not easy to collect or to have access.

As an initiative to provide source data to child's language research for language acquisition, the Child Language Data Exchange System was started around 1980's in Carnegie Mellon University. Continuously growing since, this project has incorporated not only computational tools, but several freely open language databases, covering material for analysis of second language acquisition and language impairments [9]. As part of the latter segment, the TalkBank hosts a collection for language in traumatic brain injury, i.e. the TBIBank which was worked in collaboration with the University of Sydney. The corpus was conceived as a cohort work for the recovery stage with the participation of 42 people in average per time point. It consists of seven exercises mainly narrative, from which Cinderella retelling task is mostly worked here [10, 8].

Dealing with scarce data, appropriate methods have to be taken into account to carry out significant analyses on available data, given that big-data techniques to reveal patterns are not advisable. Moreover, there are not immediate data sources, as social networks, from which we can collect additional data. Besides, even when economic and infrastructure resources have been solved, there are other issues to attend to. For example, the specialist must encourage an actual and constant disposition of injured people to contribute to the study's tasks.

Within this context, standard learning methods were used to evaluate selected features characterizing the altered language [11]. Not all learning methods were appropriate to examine limited data to have reliable conclusions. During the process, it was also observed that relations among features are complex and inherent to the subject of study with intricate connections when heterogeneous attributes are examined.

Particularly, it was observed that the correlation among characteristics varies in consecutive time

[1] "TBI is defined as an alteration in brain function, or other evidence of brain pathology, caused by an external force" [1].

[2] In this document, atypical language refers to the language that has been affected in some manner.

points, something fundamental in recovery period studies. Also, it has been noticed that varying correlation is strongly related to the stability in selected features, which has not been considered at all when selection is done [12]. Furthermore, it seems that variation in the relation among features can be directly related to the unsteady proximity that attributes present. Thus, it is convenient to elaborate on explicit information that accompanies the development of the features' values through time. Thus, there is a need to estimate internal proximity among the clustered variables and the adjustments in external relations among those clusters along a period.

The aim is to formulate a technique to trace variation in proximity of the atypical linguistic features, examining inside their groups and among groups as entities. For that, we propose to work on this problem integrating alternative methods to enquire into the development of features of the altered language after a traumatic brain injury, modeling them as trajectories in a space, tracking proximity variation in their relation along a correlation evaluation through the recovery stage.

With this, we expect to contribute to Computational Linguistics in the examination of features stability regarding their relation, and possibly to computational movement analysis[3], determining a method for tracing variability behavior between their trajectories. Furthermore, testing the proposed technique on the atypical linguistic variables of interest, we could also contribute new knowledge to language pathologists, hopefully useful in their plan treatment.

1.1 Motivation

As previously mentioned, cases of head injuries continue and increase worldwide. These numbers can be widen with some cases of SARSCOV-II, where critical episodes with oxygen deprivation can leave sequelae in language of the type of a TBI [14]. Further studies evidenced that patients who did not require hospitalization, given that they were overcoming moderate illness symptoms with no signs of neurological difficulties, also presented cognitive-linguistic problems, which was an unexpected result [15]. Other work [16] concluded that some patients require speech and language treatment after being affected by Covid-19. More recently, an analysis about local monitoring of the problems in cognitive-communication functions coming from this disease reports that about 47% of the sample presented at least mild language affectations [17].

As we can notice, the worldwide annual estimation of head traumas is growing, placing it as a leading cause of disease, disabilities, or a condition of special needs. In this estimation, the numbers coming from wrong diagnoses, or a lack of medical attention, are not considered. Following a head trauma, different types of restructured language can arise, a situation that requires a proper inter-field intervention. Hence, computational analyses can additionally produce pertinent examinations and reach reliable findings that might contribute to clinical analyses decisions.

Furthermore, the works taking into account relevance and stability in the feature selection process are scarce. Also, few examinations of modified language cover a period, and even less analyze the recovery stage. All these elements produce a setting of scarce information to understand language development after a head injury.

[3] A term recently coined [13]

1.2 Justification

The main reason to have a condition of limited data is our object of study and the context where it develops. Sensitive information management, or a lack of an adequate standard methodology to register language samples are a couple of aspects of such context. But after that, we face further computational constraints to obtain additional data. Data augmentation is not an option, since those techniques alter original instances, introducing in some cases noise to create new instances, so biasing the analysis beyond language alteration.

Dealing with limited data, alternative methods are considered to elaborate a technique to track relation variability among features. Previous work started to attend the stability issue in feature selection but continued prioritizing other aspects, such as dimension reduction or performance in a categorization task [18, 19]. Few research works inspect relation among features to incorporate in the feature filtering process, and practically no work studies feature relation variability to understand their contribution variation to define language disorders.

The use of techniques designed for large data of previous works can hardly produce reliable results. Furthermore, scarce studies have traced linguistic feature behavior during a period. So, there is a need for an approach dealing with available data and contributing comprehensible results to the related fields, a demand of physicians and medical specialists largely postponed. In particular, the attention to the disregarded problem of recognizing the development of the adaptations in the features of acquired language disorders.

In Computational Science, with an alternative approach implementing heterogeneous methods, we seek to track the variation in the relation among linguistic characteristics and also monitor the likely changes in their contribution over a period, that is, to reveal evolution modification of the features over time. We expect to achieve a technique that is explainable and comprehensible to the areas of study related to language analysis affected by head trauma. In particular, the pathologist could consider the expected results to design and implement improved recovery plan for patients. Furthermore, a technique for tracking down in time the variation among some set of attributes of an early defined relation can be adapted to another studies, for instance, in the analyses of language reorganization for the initial diagnosis of dementia.

So, the main research challenges to face are: how to extract the latent information in language disorders from the limited available data; how to establish existing relationships in the proposed context, such as the relation variability among features and their effect on the characterization of the language of study; how to evaluate the efficacy of the proposed representation to delimit narrative profile from the study cases.

1.3 Problem Statement

The study of altered language after a head trauma presents a set of challenges, such as the collected samples are worked on a project basis, so only a part of transcripts are currently available, and the sample type can not be obtained from immediate resources such as social networks. Furthermore, augmentation techniques do not apply to our context, where we examine language alterations. All these together lead us to limited data. Previous studies identifying relevant linguistic attributes and classifying these alterations showed some limitations in applying learning methods.

Specifically, when evaluating consecutive samples in two recovery stages in the analysis of the characterization of language with alterations after a TBI, we noticed a variation in the relation between the attributes, a fact that can be related to an unstable or incorrect language characterization. Thus, the problem is to track that variation in the relationship between features and how this reflects in their contribution defining the language of interest in the recovery period. In addition, we expect to extract latent knowledge of the affected language evolution and to reach explainable and reliable conclusions.

1.4 Research Questions

1. What is the appropriate measure of proximity in a topological space [4] for feature evolution comparison?
2. Once the proximity measure has been selected, how to determine the criteria for feature clustering?
3. Once the features are clustered, how do the relationships vary within each group and how do they change between the defined feature clusters?
4. What is the range of variation of the relationship among the attributes in the recovery stage, and how drastic are the changes?
5. What is the level of dependence between the variation of the relationship among features and the variability in their contribution to characterize the alterations in language?
6. What is the effectiveness of tracking the evolution of the internal trajectory of the narrative samples, as a basis for identifying the affected language instances?

1.5 Hypothesis

We hypothesize that we can trace variability in the relationship among linguistic features altered by a traumatic brain injury, through the internal development of narrative instances along the recovery stage, which can help to elucidate variation in the contribution of the evaluated features, in addition to uncover knowledge about the language development given limited data while supporting explainable conclusions.

1.6 Objectives

1.6.1 Main Objective

Formulate a technique to track the variability of the relationship among linguistic features during the recovery period, within sets or between them as entities, in language discrimination after head injury, that could shed light on the modification of the affected language.

[4] "A space which has an associated family of subsets that constitute a topology. The relationships between members of the space are mathematically analogous to those between points in ordinary two- and three-dimensional space" - Oxford Dictionary.

1.6.2 Specific Goals

1. Determine the appropriate approach and measurements for the adequate clustering of the attributes.

   The metrics to work later on features sets will be obtained.

2. Measure and identify proximity behavior between evolving trajectories of atypical linguistic features in their space.

   Initial parameters to support the identification of the attributes relation will be determined.

3. Examine relationship variability among features and their contribution through their trajectories in the evolution of alterations in language.

   Variability values to approximate likely relationships will be computed.

4. Define a procedure to evaluate the evolution of proximity between trajectories of features in their space.

   Method and metrics to trace relationship variability among features will be developed.

5. Formulate a technique to trace the internal evolution of the proximity among feature sets and the changes in their vicinity as entities.

   A complete procedure to examine the relationship and contribution variability among atypical linguistic features will be achieved.

6. Test and evaluate the proposed approach.

   An evaluation to discern possible adjustments to our proposed technique will be done.

1.7 Scope and Limitations

The scopes ($*$) and limitations ($\otimes$) are summarized below:

- $*$ The research focuses on identifying the variation in a given relationship among linguistic features extracted from transcripts of samples from patients after head trauma for the study of language sequelae.
- $*$ The study covers the recovery stage, a period of more sensitive adjustments of the examined language.
- $*$ Narrative instances are studied, a genre indicated as appropriate for analyzing language disorders. Mainly the task of the Cinderella story retelling is examined. Specialists previously worked on the transcription of the samples.
- $*$ The analysis incorporates study and control samples available at the same site, though from different projects.

⊗ The results and conclusions will be based on the analysis carried out on the available samples (limited data).

⊗ The type of structure of the used samples will be necessary to replicate the study in another data set.

⊗ The control set is from a unique sample.

1.8 Expected Contributions

We foresee the following contributions from this proposal.

- In the field of NLP focused on language disorders analysis, a process that tracks relationship variation among linguistic features of language disorders after head trauma, that could be adapted to other studies requiring an exploration of this type (e.g. the early diagnosis of Alzheimer).
- In computational linguistic analysis, a process to follow the evolution of the patient's language disorders developing a narrative task.
- In NLP in general, an alternative approach for the analysis of limited data.
- In clinical language treatment, hopefully new knowledge about the internal evolution of the patients' narrative samples.

2 Background

The TBI Corpus comes from a more extensive project, a longitudinal study in which the University of Sidney collaborated with Carnegie Mellon University. It consists of five samples, four at periodic time points every three months and one more at twenty-four months after a head injury. Pointing out that narrative tasks are recommended to assess language restructuring, this corpus collected mainly narrative exercises, as generative based on pictures or storytelling based on a book.

Original samples were recorded in audio or video, and professionals worked on the transcripts. The control set comes from another project, also available on the site. It has a similar number of participants but with a single sample. In general, the project implemented standardized processes in an effort to maintain conditions to minimize the bias on the results of extracted analyses. For instance, a pair of technicians transcribed each sample independently with a statistical agreement above 90%.

As mentioned above, the samples in the Corpus consist mainly of narrative instances since this literary genre has been suggested [20] as a task that can capture alterations in language. Thus, these types of cases are appropriate in studies of language disorders and, in our case, the sequelae of head trauma. The narrative exercise is indicated when analyzing the cohesive mechanisms as they occur in the production of everyday discourse. Also, a structured discourse has been recommended for comparing samples over time [21].

On another line of research, computational movement analysis, an interdisciplinary research field capturing and structuring data studies, handles the proximity conception in a space. Also, it examines the movement phenomena considering an abstract space. Besides the geographical context, it looks to advance the understanding of the processes operating it. This field approaches data drawing on methods from temporal analysis in one space, computational geometry, and statistics, among other disciplines that enable it to structure low-level movement data to extract high-level process understanding [13].

In addition, regarding the type of nature-inspired methods coming from biology, physics, and chemistry phenomena observation, particle physics indicates that when analyzing the behavior of a set of particles as a whole, the interaction and their operating forces must be heed [22]. When analyzing these lines of investigation, we establish a link between the existing connection in the logic of the narrative development (from the linguistic perspective) and a balance of forces in the form of particle vectors (from Physics), which provides a basis for the proposal.

We introduce the analysis of the variability of the relationship among features, working on the evolution within the narrative instance expressed as trajectories in its space. Based on this analysis, we define a measure of proximity between trajectories with alternative methods as part of the proposal developed later in section 4. Next, additional details about the narrative task and the physics of particles are provided. Definitions of the measures employed in the comparison of trajectories are also included.

2.1 Narrative task

Understanding a narrative as the telling of a real or fictional story, a narration has a structure. It has to follow certain coherence to make sense throughout the narration. Thus, the fragments that form the tale are interrelated. Evaluating chunks per participant's intervention, we expect that the graded responses are substantially connected. Thus, to analyze how relations, such as correlation, vary among the set of studied features, we process the narrative sample at different time points inside out. To support this, we require to go through narrative conceptualization. Though there is no single direction to comprehend the structure of narration [23], we recap a few annotations to delineate our angle to make sense of our approach. Authors of the work [24] state Narrativity as "the degree of narrative organization of a system" after observing that "the basis on which narrative is perceived/defined is a network of dynamic relations". Different relations are identified inside the minimal definition of narrativity that Piper et al. [23] delineate as (i) a co-determinative relationship between the act and the physical universe, a real or fictional setting, just as between the agents and their actions, (ii) the distribution of the events outlines the relational construct besides intelligibility, (iii) the temporal, logical, or causal relations among narrative sentences.

A marker of weak relations might be the discrepancies between the story and the discourse, which depict anachrony. Narrating accounts for an expected sensitivity that gives origin to language variation in a narration. Psychology understands a narrative as "a latent organizing principle of human reasoning" [23]. In the current research, we integrate sets of splits by time alignment for the narrative instance, as referred to in [23], and calculate variables at those time points, our group of particles to analyze.

2.2 System of Particles

As described above, we observe dependent relationships among the elements of a narrative, thus among the variables that evaluate those components forming the narration. We have a set of indexes measured at different points in time along our narrative samples, where narratives do not have physical but abstract movement in the sense that those markers present modifications at the regarded period. Additionally, the work [25] notes that "you cannot generate a thought without some electrical signal moving in your brain".

Thus, we approximate the level of dependence among features based on their distribution along a time interval. For this, we express the group of indexes of a narration instance as a system of particles, where a system including multiple particles must account for the inter-particle forces and the inferred bulk motion of the entire system [22], as defined in Physics. After generating trajectories from attributes, we compute the center of triangle built by segments to obtain a representative value of the forces interacting in that segment along specific midpoints to generate new trajectories. This approach is contrasted by two measures suggested in the literature [26, 27] to analyze trajectories.

2.3 Measures

In this section, we explain the metrics for the proposed technique, and then we develop in more detail the particular used measures for the different methods and processes that compose our approach.

2.3.1 Metrics for assessment

The proposed technique implies the release of a model aimed to address limited data that will be evaluated with area under the curve (AUC), assessing its capacity of categorization, and if time allows, will be compared against generalized mixed models (GMM) on how well the fixed effect is represented by the random effects.

2.3.2 Particular used measures

In general terms, we give the notions of Hausdorff and Fréchet metrics, for formal and detailed treatments see [28, 29] for the former and [30] for the latter.

Hausdorff measure Informally, the survey [31] explains the Hausdorff distance as "the maximum of the distances between each point in a set Q to the nearest point in the reference set $(trajectoryT) : d_{Hau}(Q,T) = max_{q \in Q}\ min_{p \in T}\ d(p,q)$".

Belogay *et al.* [32] define the concept as follows. Given A and B bounded sets in Euclidean space (E^m), with:

$$d_B(A) = \sup_{a \in A} d(a, B)$$

Defining the distance from point a to set B (the $minB$ calculation) as:

$$d(a, B) = \underset{b \in B}{inf}\, d(a, b)$$

In general $d_B(A) \neq d_A(B)$ and the $\text{Ball}_b(\varepsilon)$ is the closed ball of radius ε centered at point b. The *Minkowski* $\varepsilon - sausage$ of B (B_ε) is:

$$B_\varepsilon = \bigcup_{b \in B} \text{Ball}_b(\varepsilon)$$

So, $d_B(A)$ is equal to the smallest ε, such that, A is contained in the $\varepsilon - sausage$ of B, where A, B can be the graphs of curves. Then, in the bi-dimensional space, $A = \{(x(t), y(t)) \mid t \in [0, 1]\}$, for some parameterized curve γ. The Hausdorff distance between A and B will be taken as:

$$h(A, B) = d_B(A) + d_A(B)$$

Fréchet measure. The setting is a walker bringing a dog with a leash, each generating a path, and the idea is to have the minimum length of the leash during the unidirectional trajectories course of the dog and its walker [33, 27].

Pearson. The correlation coefficient ρ, measures the linear association between X and Y, the ratio of the covariance between X and Y to the product of their standard deviations, i.e.,

$$\rho = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

where μ_X, σ_X, and μ_Y, σ_Y are the respective means and standard deviations of X and Y, with $-1 \leq \rho \leq 1$. Besides, $\rho = \pm 1 \iff Y$ is a linear function of X, and $\rho > (<) 0$ means a positive (negative) linear relationship between X and Y. If Y independent of $X \Rightarrow \rho = 0$, and $\rho \neq 0 \Rightarrow X$ and Y are dependent. The estimate is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Assuming that the random vector (X, Y) has a bivariate normal distribution, then the estimator r is the maximum likelihood estimate (MLE) of ρ, [34].

Kendall. Given two pairs of random variables, $(X_1, Y_1), (X_2, Y_2)$, independent random vectors with the same distribution as (X, Y), a jointly continuos random vector. Kendall's τ_K measures monotonicity between X and Y in a probabilistic sense, defined as:

$$\tau_K = P[sign\{(X_1 - X_2)(Y_1 - Y_2)\} = 1] - P[sign\{(X_1 - X_2)(Y_1 - Y_2)\} = -1]$$

where $-1 \leq \tau_K \leq 1$.
$\tau_K > 0$ indicates increasing monotonicity, $\tau_K < 0$ indicates decreasing monotonicity, and $\tau_K = 0$ neither monotonicity. If X and Y are independent then $\tau_K = 0$, with the contrapositive true, i.e., $\tau_K \neq 0 \Rightarrow X$ and Y are dependent.
(X_1, Y_1) and (X_2, Y_2) are concordant if $sign\{(X_1 - X_2)(Y_1 - Y_2)\} = 1$, and discordant if the sign is negative. Given a random sample $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$, counting the number of concordant pairs and, substracting the number of discordant pairs, the standardized estimate of τ_K is:

$$\hat{\tau_K} = \binom{n}{2}^{-1} \sum_{i<j} sign\{(X_i - X_j)(Y_i - Y_j)\}$$

$\hat{\tau_K}$ is distribution-free, with μ 0 and σ $2(2n + 5)/[9n(n - 1)]$, [34] .

Spearman. Consider the random sample $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ with $R(X_i)$ the rank of X_i among $X_1, X_2, ..., X_n$, and likewise $R(Y_i)$. The estimate of ρ_S is the sample correlation coefficient with X_i and Y_i replaced respectively by $R(X_i)$ and $R(Y_i)$, it estimates monotonicity between the samples. Let r_S denote this correlation coefficient.

$$r_S = \frac{\sum_{i=1}^{n} (R(X_i) - [(n+1)/2])(R(Y_i) - [(n+1)/2])}{n(n^2 - 1)/12}$$

r_S varies between ± 1. With a strictly increasing (decreasing) relation between X_i and Y_i when $= \pm 1$. If X and Y are independent, it follows that r_S is a distribution-free statistic with $\mu = 0$ and $\sigma = (n - 1)^{-1}$, [34].

3 Related work and State-of-the-art

The study in [35] focused on the stability of algorithms selecting features motivated by the fact that with a sharply growing dataset size, most of the efforts centered on the number of features, and the stability in the selected set was practically unattended. After applying some changes to the determined training sets, the work examines different domains for effect on classification tasks based on ranking measures (including information gain or RELIEF), or learning algorithms based on supporting vector machine. The authors found no steady feature selection method for the several tested settings, and that there are signs of dependency between the problem and the approach. Additionally, stability property was not reached with the top-ranked feature group evaluated for classification error, noticing that high instability does not necessarily imply low classification capacity. They attributed this to an existing redundancy in the set but with no experiments to show it. Likely, this work is a pioneer in heeding the stability property of algorithms selecting features.

Bioinformatics [36] started stability examination attending its role in feature selection when working DNA microarray data, assuming unstable feature subsets originated in an imprecise selection. The authors added that there is little work on the optimum parameters using experimental perturbations (randomly removing instances from the original datasets), partitions, or the number of folds in cross validation. Also, the review collected several similarity measures, such as that of Kalousis, Hamming distance, Kuncheva, or the ASM (adjusted stability measure), though observed that theoretical assumptions to be satisfied are unclear. Among the causes of instability, they identified a disregarded stability analysis, multiple true markers highly correlated, or variance in the data. Additionally, they recognized group and ensemble feature selection as the worked perspective aiming to alleviate the instability problem. The researchers concluded that the work in feature selection stability is scarce or null. For instance, a pending task is the comparison of the proposed measures to evaluate it or which works best in a given framework.

The study [12] notes that the stability analysis work is incomplete since efforts are mainly invested in seeking a stable reduction in the number of features. The authors observed that the direction of the examination should be oriented to maintain stability in the contribution of the features, characterizing the subject of study when they are worked by some selection criteria. In particular, clarifying that the notion of stability from the perspective of particle swarm optimization (PSO) does not have this approach. To work on this deficiency, they propose a variant of the BPSO binary algorithm called COMBPSO. This adaptation involves a linear combination of three modeled concepts. So, they bring together the concepts of relevance, consistency, and classification performance of the feature subset to define its fitness function.

In addition, the researchers introduced a turbulence operator to avoid a premature convergence of their fitness function. Furthermore, they established inertial weights based on a sigmoid function. Also, working on an asymmetric interval for the runs, they suggest empirical values for different parameters. The basis was that feature consistency on which the subset consistency is based helps to calculate the amount of overlap between the selected subsets to observe their stability. With this, researchers achieved comparable scores between BPSO and COMBPSO and smaller subsets with lower error in the discriminating task.

Afterward, aiming to obtain the minimum and effective subset of appropriate features, the work

in [18] presents a new selection algorithm rooted in feature stability and correlations, looking for high accuracy and dimensional reduction, with priority of the former over the latter. The preprocessing engine works on dimensional reduction establishing experimental thresholds for stability and correlation. The algorithm consists of two stages, based on the thresholds, defines a group of models later ranked by accuracy, however they noticed that optimum reduction rates do not necessarily mean acceptable levels of accuracy. Comparable results with the state-of-the-art are achieved. But the recurrent observation in previous works that, 'among the existing feature selection methods, none works well for every dataset' remains after discussing their results.

Noticing that selected features are evaluated for relevance for the task at hand, underlining the importance of stability property, several stability measures are identified based on index, ranks, or weights. However, none complete the properties observed for feature selection stability measures, defined with upper and lower bounds, maximum, correlation for chance, and monotonicity. In addition to a feature information strategy, the evaluation of the connection between features is observed as an aspect to contribute to the solution of the instability problem further to sample weighting or parameter optimization, and the authors concluded that progressive research in this line is a need [37].

Besides, additional information may be required regarding a neutral or noise function, depending on the focus of the examination. To differentiate these roles in addition to the relevancy or contribution, one of the first steps in the feature selection process is clustering[5]. In this aspect, a simple clustering conception based on an ant colony when they clean their nest was introduced in 2001. Ants and n-dimensional vectors (dead bodies) share a matrix of MxM where ants can pick up or drop out the objects from the grid cells according to a probability rule that determines the similarity of the object with those around its [38]. In 2006, a graph was worked to solve clustering, now collecting a predefined number of documents to form clusters from graph nodes. There, the Euclidean distance determines the proximity among documents. This same technique was employed later (2011) to solve a fuzzy clustering problem [38].

The krill herd algorithm, also observed to behave well for data clustering, has been working together with the harmony search algorithm that mimics the improvisation process in music, appropriate for optimization of numerical functions and clustering of text and data. This composition was proposed as an improved strategy for data clustering task to variables with continuous value, identified as a complicated process. The integration of the global exploration operator of the harmony search technique alleviates the problem of premature convergence of the original krill herd algorithm [39].

Perspectives previously summarized are worked at one discrete time window. In the computational context, within movement analysis area, is possible to examine proximity variation of movement in its space taking into account temporality [40]. According to [41], a trajectory models an entity in movement. Trajectories produced by the examined variables can be contrasted by criteria such as similarity, pattern detection, or clustering. Besides, *group* identification accounting patterns are based on proximity in a space, while clustering is based on similarity assessments. Starting with the idea of a group as a set of entities that hold along a defined period, a few other requirements can be added to tune properly a structure, thus work it as a group of trajectories [41].

[5]Congregate elements of a set in subsets given a criteria.

In the movement analysis field, the work [40] applies techniques in a space to identify collective trajectory patterns. To achieve this, measurements used for individual trajectories are extended to measure them in group. In general, the author treated the groups under three views. i) The representative, where a trajectory summarizes the information of the evaluated trajectory set, ii) the complete, in this view, each trajectory is considered for the measurements and finally, iii) by area, defining a circle radio to go through each trajectory and the union of the areas is regarded as measure, introducing some new group measures, such as density. One of the main contributions is the refined group concept of moving entities, previously introduced by Buchin and, defined based on distance, duration, and size. Subtle adjustments to distance conditions are proposed, and the corresponding algorithms adaptations were designed in consequence.

Regarding the feature analysis process, other aspect to consider is feature selection. Again, on the line of approaches based on nature, the Reptile search algorithm (RSA) along with the Remora optimization algorithm (ROA) integrate the method HRSA, in which a novel shift method to drive transition mechanisms is added. This procedure is called the mean transition mechanism MTM, and is used to control the searching process and the transition mechanism between RSA and MT. The principle of the MTM is to regulate the search when there are no gains from the fitness function after five iterations. With this fusion, authors handle optimal local problems alongside the unbalance between exploration and exploitation mechanisms, with promising results for most of the tested contexts [42].

From linguistic perspective, the *narrative* is the basis of analysis, an element of discourse that is the indicated Gold Standard to approach language alterations as an effect of TBI [43, 44]. The narrative exercise is a fundamental task for language examination [45], such as identifying strengths and weaknesses in therapy direction. But beyond that potential clinical use, a narrative-based intervention might support treatment to improve quotidian communication capacity. Though unexpectedly, the authors found few research works aiming at it [46]. Some topics of the analyses are story grammar elements [43, 47, 48], main concept analysis (MCA) [49], words per narrative, T-units (terminal units[6]), and C-units (communicative units[7]). Here, a pertinent observation is that narrative discourse intervention reported mixed results [46] on six identified works aimed at understanding the structure and elements of narratives, recommending further research to recognize the best discourse task stimuli for assessment and treatment. Additionally, robust methodological designs should be incorporated in experiments, though there is minimal guidance to direct research efforts seeking evidence to support clinical practice on language effects from TBI.

From another angle of our proposal, we have that the complexity of the object of study leads to scarce data, and data augmentation that could be the direct option does not work in any context. A recourse of data augmentation is the transformation of the original samples to produce more instances. The modifications can include the addition of random noise, masking, deletion, swapping or replacement of words, or even the reorganization of sentences or chunks. These two last observed with a negative effect on the sample, producing instability [50, 51]. Considering that we are examining language alterations produced by head trauma and the observation that the dataset and the setting regulate the best data augmentation principle [51], these techniques are not

[6]T-unit consists of a main clause and all subordinate clauses associated with it, [48].

[7]"An independent clause along with any of its attached dependent clauses", [48].

recommendable in our current context.

NLP approaches change rapidly, although they moved predominantly to designing architectures fed with huge data [31, 52]. In addition to the observation that the border definition of low-resource is unclear, the data dependency premises must be taken into account. Since, often such premises are not met when techniques are adapted to other contexts, to attend domains with limited data. Also, the suitability of pre-trained models (e.g., mBERT) for settings that naturally belong to low-resource techniques is arguable [53]. In recent years, evidence that non-neural methods achieved better results for some NLP tasks than neural techniques was shown, were these last only started to improve with increased samples [31, 53, 54, 55]. We close this part by emphasizing the call to prevent ordnary language assessment to follow-up language disorders, as noted for infants, given their limitations in evaluating non-typical language [56].

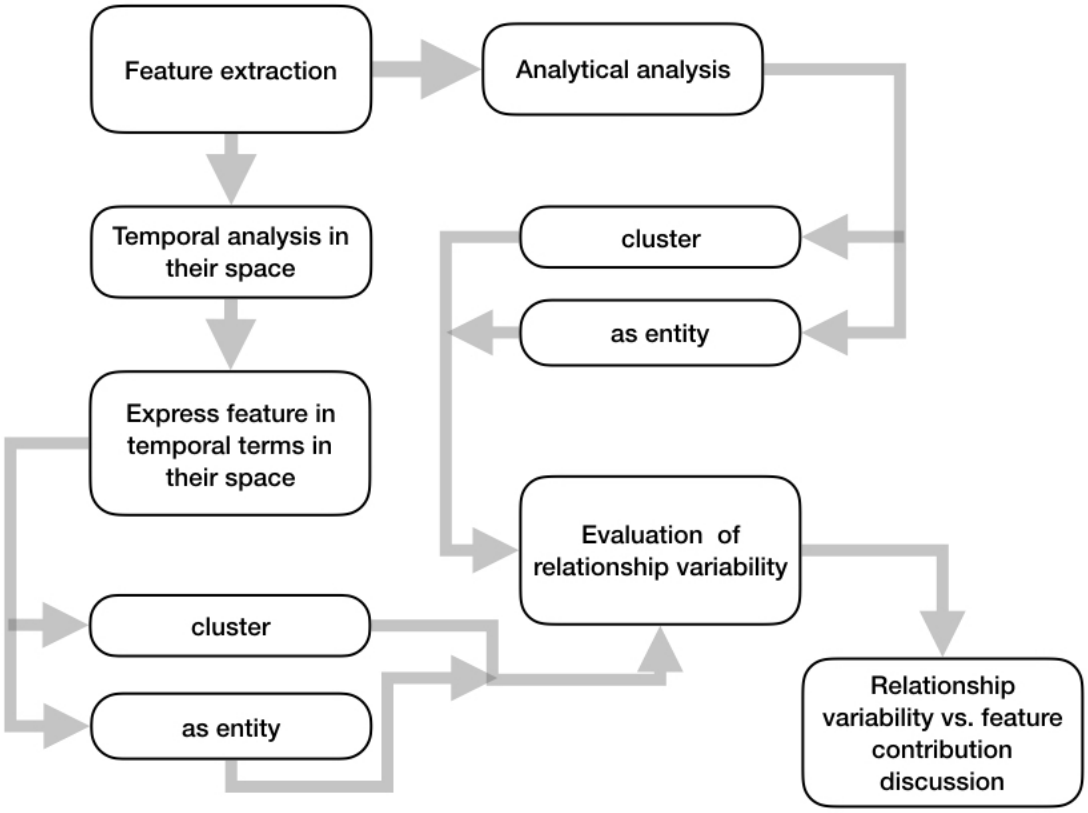


Figure 4.1. Research proposal diagram

4 Research Proposal

Once the research problem 1.3 has been stated and the general 1.6.1 and specific 1.6.2 objectives have been established, we proceed to detail the methodology that will allow us to achieve them.

After extracting the features per participant, on one hand they are evaluated analytically and on the other, their evolution is followed in time, in their corresponding space. For the second part, the data is processed to express it in temporal terms in its space, which implies generating sub-samples based on the time alignment of the original sample.

Subsequently, the variability of the relationship among features within the defined sets and between them will be evaluated, to then examine their contribution to the discrimination of alterations. These steps are illustrated in figure 4.1.

4.1 Methodology

The methodology summarized in steps and elaborated immediately below.

1. Determine the feature set to carry out the analysis.
2. Extract and preprocess selected features.
3. Identify the measures of comparison between characteristics to use.
4. Determine the feature clustering criteria.
5. Generate the corresponding representations and values to evaluate.
6. Evaluate the variability of the feature relationship.
7. Adjust previous processes and procedures to improve results.

Determine the feature set to carry out the analysis. Given that this was part of previous research, in this step it was determined to consider lexical, syntactic and pragmatic features, for this research. The lexico-syntactic set is defined for a narrative language analysis[8] consisting of little more than 40 characteristics, of which 25 are lexical, basically frequencies of parts-of-speech. Among these lexical features we find: total number of words, verbs, nouns and proportion of nouns by verbs. The syntactic part includes inflections such as comparatives, superlatives, irregular verbs, third persons, and progressive tense markers.

The pragmatic feature group consists of indices to assess fluency. This set, also with more than 40 features, includes frequencies, proportions and relationships to estimate, among other aspects, phonological fragmentation, the use of monosyllables or the presence of stuttering [9, 57, 58]. Examples of these are the repetition of parts of the word (PWR) or the total of typical disfluencies that consider repetitions of phrases, full or empty pauses. Figure 4.2 illustrates the described feature sets.

Extract and preprocess selected features. For the analytical part, the feature extraction will be worked by participant and task considered (the retelling of the Cinderella story). For the temporal analysis, once the files per participant per task have been generated, the alignment time in each sample will be identified and the corresponding sub-samples will be generated from which the features by type will then be extracted. Already within the analysis, one part will work the variables as they were extracted and another will require working them in the unit interval.

Identify the measures of comparison between characteristics to use. For the analytical part, the correlation based on Spearman, Pearson and Kendall will be considered. Within the temporal examination, the use of the proposed Hausdorff and Fréchet[9] similarity measures is planned. In addition, it is proposed to carry out a process to estimate the proximity between the characteristics represented as trajectories.

[8]This group is part of the standard called Northwest Narrative Language Analysis (NNLA), which is defined as part of a project for the identification of recovery patterns of a specific type of language disorder.

[9]Spearman, Pearson, Kendall, Hausdorff, and Fréchet measures were defined in section 2.3.

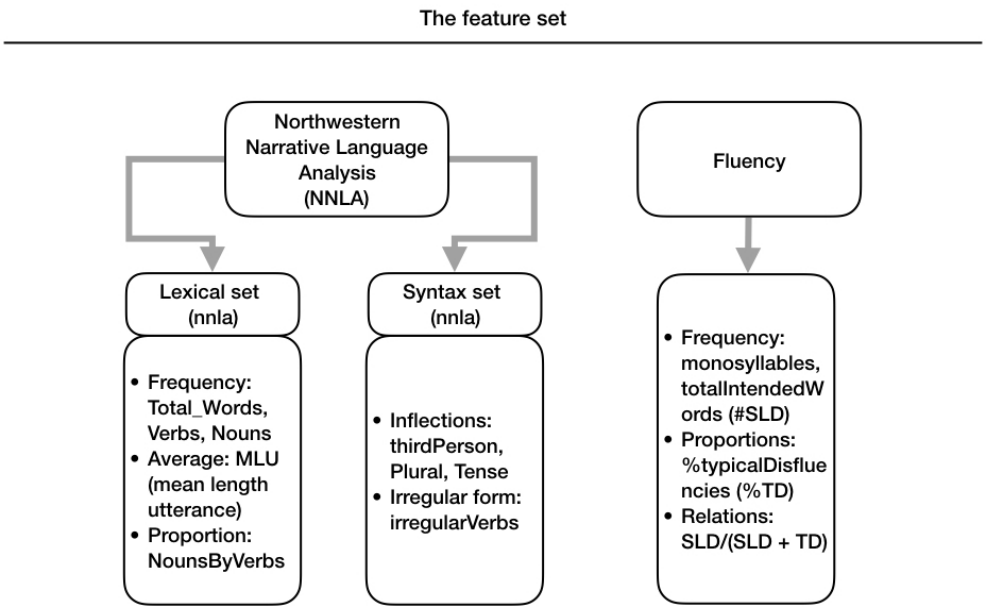


Figure 4.2. Used feature sets

Determine the feature clustering criteria. In the initial experimental part, feature groups already defined in previous studies are considered. For subsequent experiments, the identified comparison measures will be reviewed whether they are appropriate to group the worked features, or whether a new process must be proposed considering other bases for that purpose.

Generate the corresponding representations and values to evaluate. To work on the representation as a trajectory of the development of the narrative sample, the corresponding graphs will be generated. Moreover, other sets of values will be calculated, necessary to carry out the calculation process of the proximity measure, in addition to the Hausdorff and Fréchet indices and the Spearman, Pearson and Kendall estimates.

Evaluate the variability of the feature relationship. Work will be done to experimentally establish a reference variation interval for each type of feature. It seeks to have an adjustment ϵ value based on the interval, which will determine the variability of the relationship among features.

Adjust previous processes and procedures to improve results. The necessary adjustments will be made in the entire process to improve the results, such as the reduction of the variability in the relationship among the features and the discrimination capacity of the features, for the studied disorders.

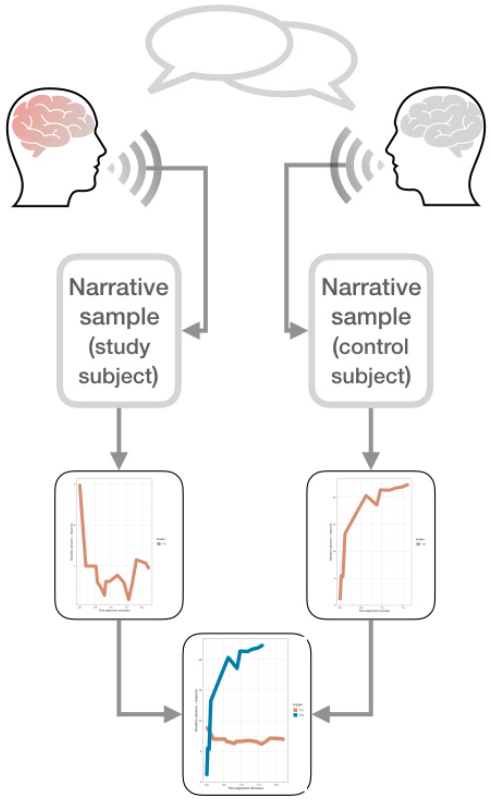


Figure 5.1. Sample extraction & comparison

5 Preliminary Results

Diagram 5.1 shows in a general way the process that we follow as part of our experiments. We start with two participants, one who has been affected by head trauma, the other is a control participant.

Different samples were obtained from both, in our context we consider the narrative instances that will be worked on to obtain a representation in the form of trajectories for which the time data is incorporated. In this first experimental stage, the samples were compared mainly for the lexical feature of *total_words*. Based on this comparison, the proximity measure that we introduce, is worked on and the values of the measures that have been indicated for the comparison between trajectories are also calculated.

The following sections detail the particular aspects worked on until the submission of this proposal.

5.1 Determine the feature set

For this research, we started by considering a set of lexical features, a couple of syntactic sets and a group of pragmatic features. These showed their discriminatory capacity in previous study,

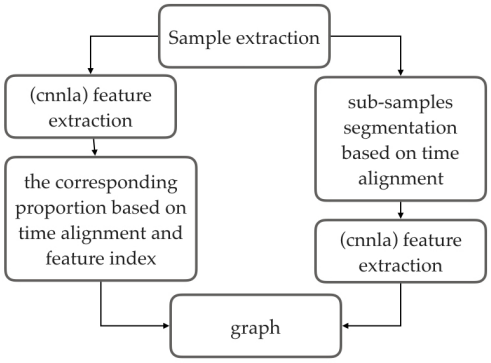


Figure 5.2. Sample segmentation steps

for the characterization of language disorders after a head injury during the recovery period. In addition, they have the property of reflecting sensitive changes in samples of subsequent periods.

Thus, we worked on generating a graph representation with one of the syntax feature sets, as part of our exploration to understand feature evolution. Triangle maps were associated with a defined context-free grammar (cfg) to trace how responses evolve. The results of this initial exploration were summarized in an article submitted and accepted at a conference, detailed on section 5.4 (No.1).

The property of this group of features of taking values in the set $\{0, 1, 2\}$ allows to generate triangle maps defined by a cfg, but limits a representation as path. So, this set is disregarded and we continue our analysis with the lexical set, the other available syntactic set, in addition to the pragmatic feature. Sets already described in section 4.1.

5.2 Extract and preprocess selected features.

Already in the feature extraction step, we look for a partial understanding of the evolution of linguistic features of the altered language through their variability through time. For this, the development of responses are tracked through sample instance during a period, to examine how those features progress through the narrative task. Thus, we need to add the time data.

5.2.1 Time integration

The exercise of completing a storytelling task implies the coordination of different cognitive abilities during a time interval. That is, the narration *evolves* through time, as it was explained in section 2.1. The group of responses on that development could add information to comprehend language alterations.

With this hypothesis, we revise and prepare part of the data to integrate time information based on the alignment that initial samples contain. This step was approached in two ways presented in figure 5.2. First, each time alignment mark was multiplied by the corresponding proportion of the total score in reference to the whole time of the worked instance. Second, original samples were partitioned to produce cumulative sub-samples, on which we extracted lexical indices to experi-

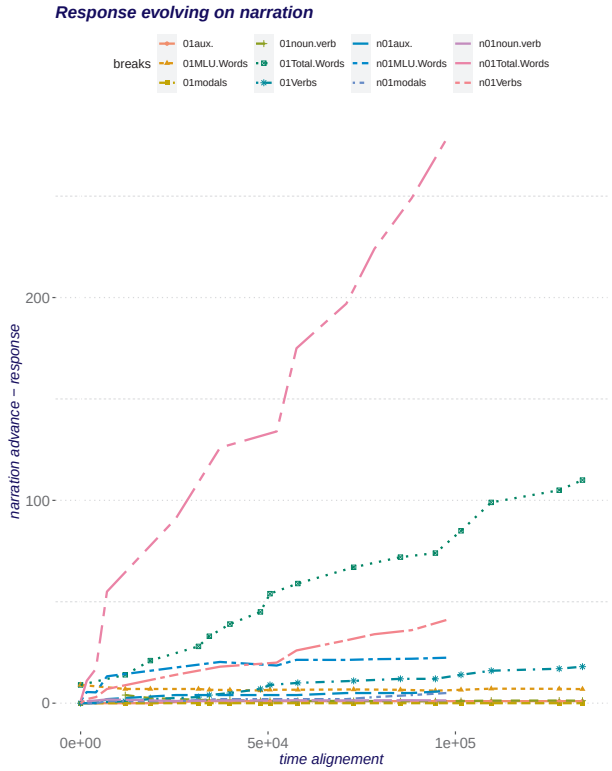


Figure 5.3. Sample data segmentation for the first sample of the control and study group.

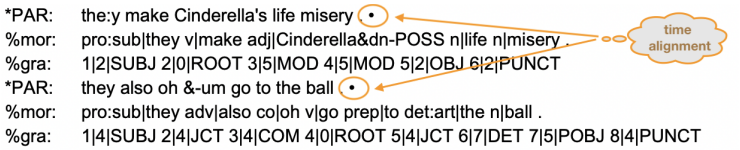


Figure 5.4. Bullets are time alignment

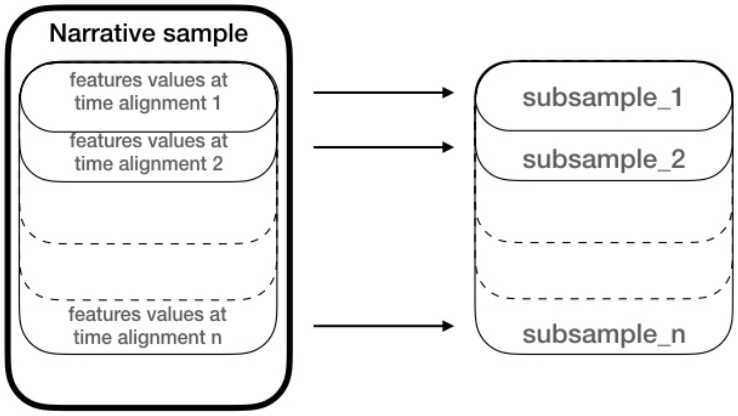


Figure 5.5. Sample segmentation based on time alignment

ment with temporal representation in their space. Figure 5.3 is a pilot example of the achieved representation, where we observe feature development on the full narration of participants. The curves correspond to a group of feature scores reached by the first subjects of the study and control groups. We notice that the participant belonging to the study group took longer to complete the exercise.

5.2.2 Sub-sample extraction

The second approach to integrate time data in the extracted information, as explained above, requires segmenting the original instance. So, we partition the sample instance in each time alignment as marked by the bullets in the source data, as shown in figure 5.4, representing the alignment of each participant's intervention.

Figure 5.5 illustrates the generation of the segmented files. Then features are extracted by subsample, to have the development of the participant's response based on its variability. The current variable along the set of its correlated complementary features will be evaluated for the first participant of study and the first control instance for the sub-samples generated based on time alignment.

5.3 Identify the measures of comparison.

As explained in previous sections, we are analyzing the evolution of the response of the participants by examining the samples as trajectories. Before addressing the comparison measures, we describe some scenarios (section 5.3.1) that we face in their generation, where for example, the ideal case is explained. Afterwards, details of the calculation of the proximity measure that we

propose (section 5.3.2) as part of the representation of linguistic characteristics as trajectory are given, and finally, we close the section with the comparison measures (section 5.3.3).

5.3.1 Scenarios in trajectories

Within the distributions of the analyzed features, in their representation as a trajectory, we identified at least three scenarios that we explain next, first we explain what we have called the ideal case.

The ideal case. We identify the ideal case when the arrangement of the vertices of the generated trajectories do not intersect, in addition to allowing a simple triangulation to be generated by alternating the vertices between both compared trajectories and completing it with *the rest*, the triangles formed with the segment of longer trajectory.

(i) In the first scenario we note that frequency features (explained in section 4.1) tend to maintain an appropriate distribution that allows the generation of trajectories that do not intersect in such a way that they fall in what we just described as the ideal case. With these characteristics we started our experiments with the latest approach, figure 5.7(c) in section 5.3.2 shows that this distribution allows an adequate segmentation in triangles, necessary for the proposed proximity measure.
(ii) In the second scenario, the set of features based on their relationships are more likely to intersect their trajectories. Thus, these cases will be tackled later in the research.
(iii) Finally, advancing in the experiments, we noticed one more situation. Although the trajectories do not intersect, the time alignments in which we extract the features can differ in length, distribution, and number, which does not allow an ideal triangulation previously described. Figure 5.6(a) is a simplified representation of this scenario, that is shown with actual data in figure 5.6(b) (MLU average).

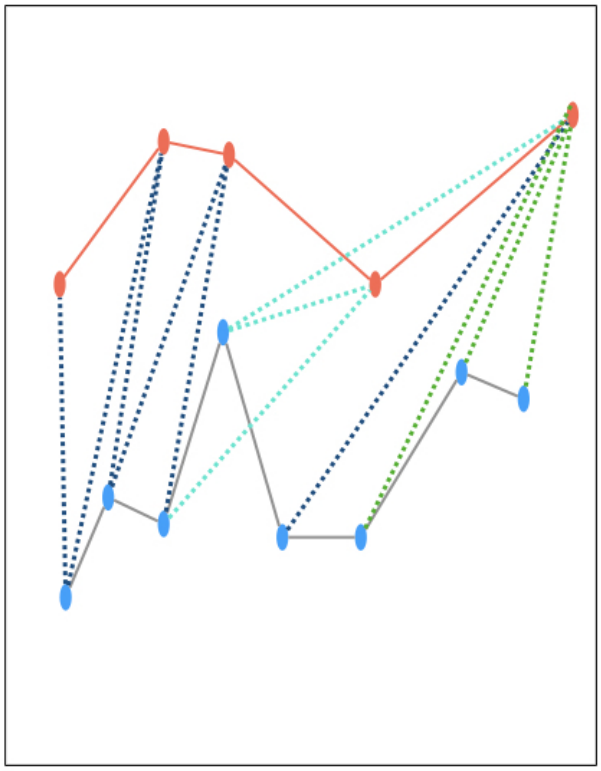
5.3.2 Proposed proximity measure

We started from different observations derived from previous works, bibliographic review and experiments to reach the proposed proximity measure that we explain below.

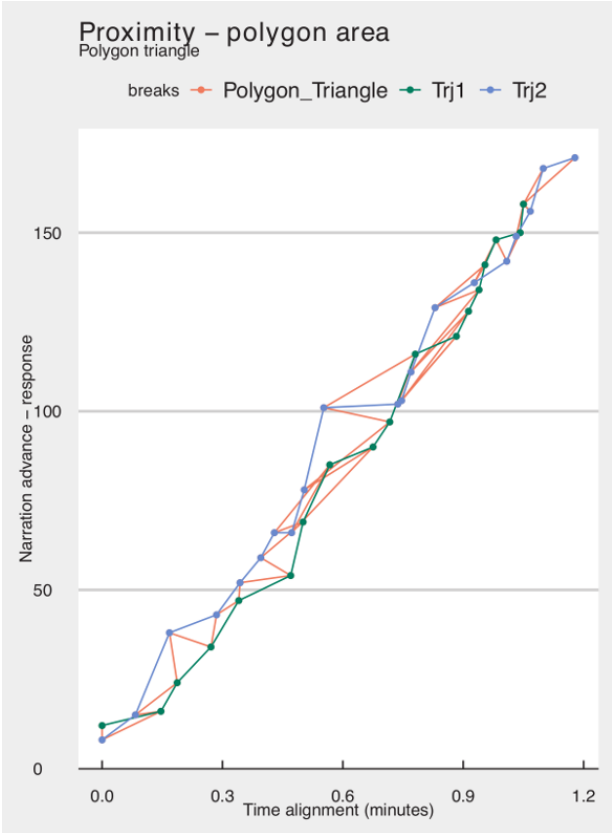
When contrasting the evolution of the response of a pair of participants through linguistic characteristics represented as a trajectory, we have to estimate how close they are. After discarding the entire area delimited by both trajectories (figure 5.7(b)) as a measure of proximity, we work on the segmentation of this area in triangles (figure 5.7(c)), calculating each of their centers, and now obtaining a new trajectory (figure 5.7(d)) from these points.

The rationale for this is that we treat each triangle as a circuit of particles, and taking the notion of center of forces of an arranged set of vectors as in Physics. Thus, we calculate the center of each segment as the point that represents the balance of forces, in the sense of connection through narration, as explained in section 2.1.

The subsequent observation revealed that the midpoints of the inner sides of the triangles appear to correctly represent the merged response of the two trajectories. This because the curve produced

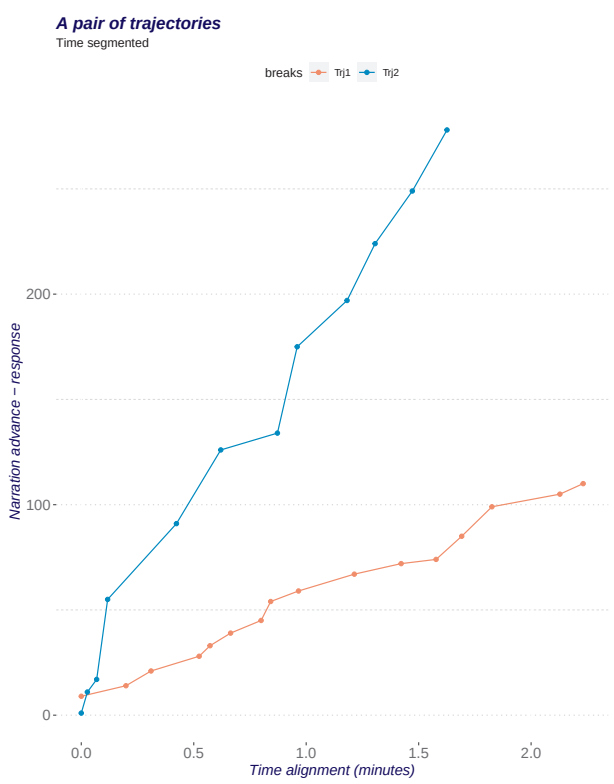


(a), Inappropriate triangle-segmentation

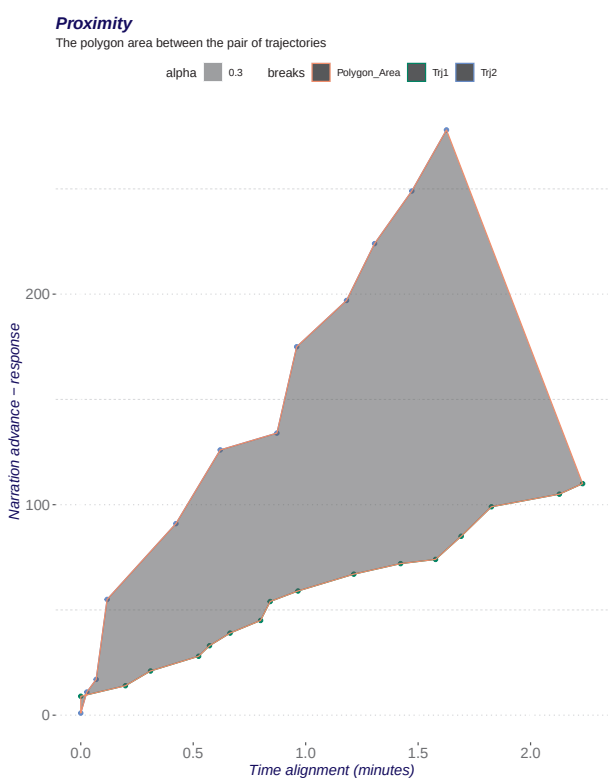


(b), Intersecting paths

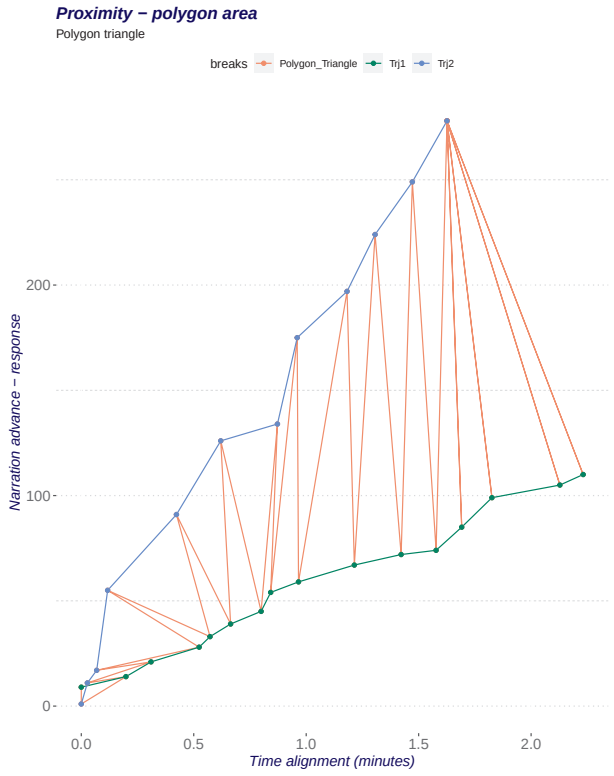
Figure 5.6. Non-intersecting trajectories - wrong triangle-segmentation



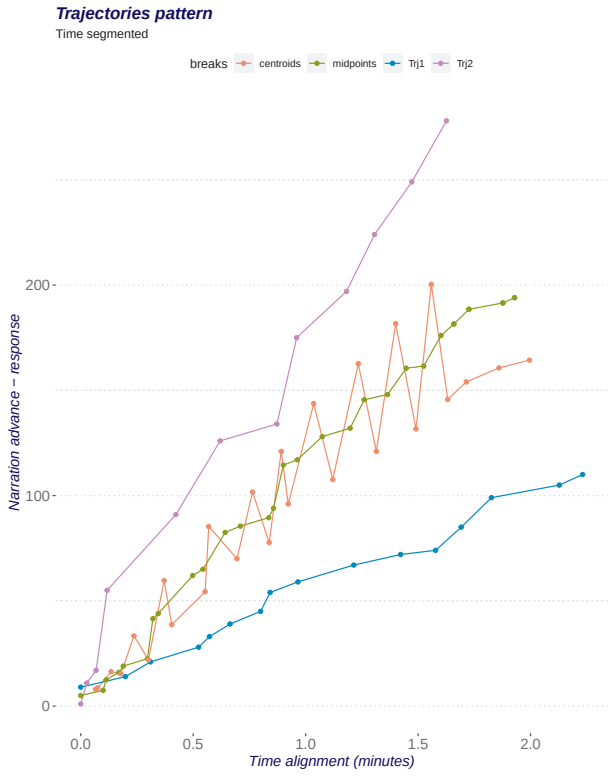
(a), Trajectories NBI01-TBI01



(b), Polygon area

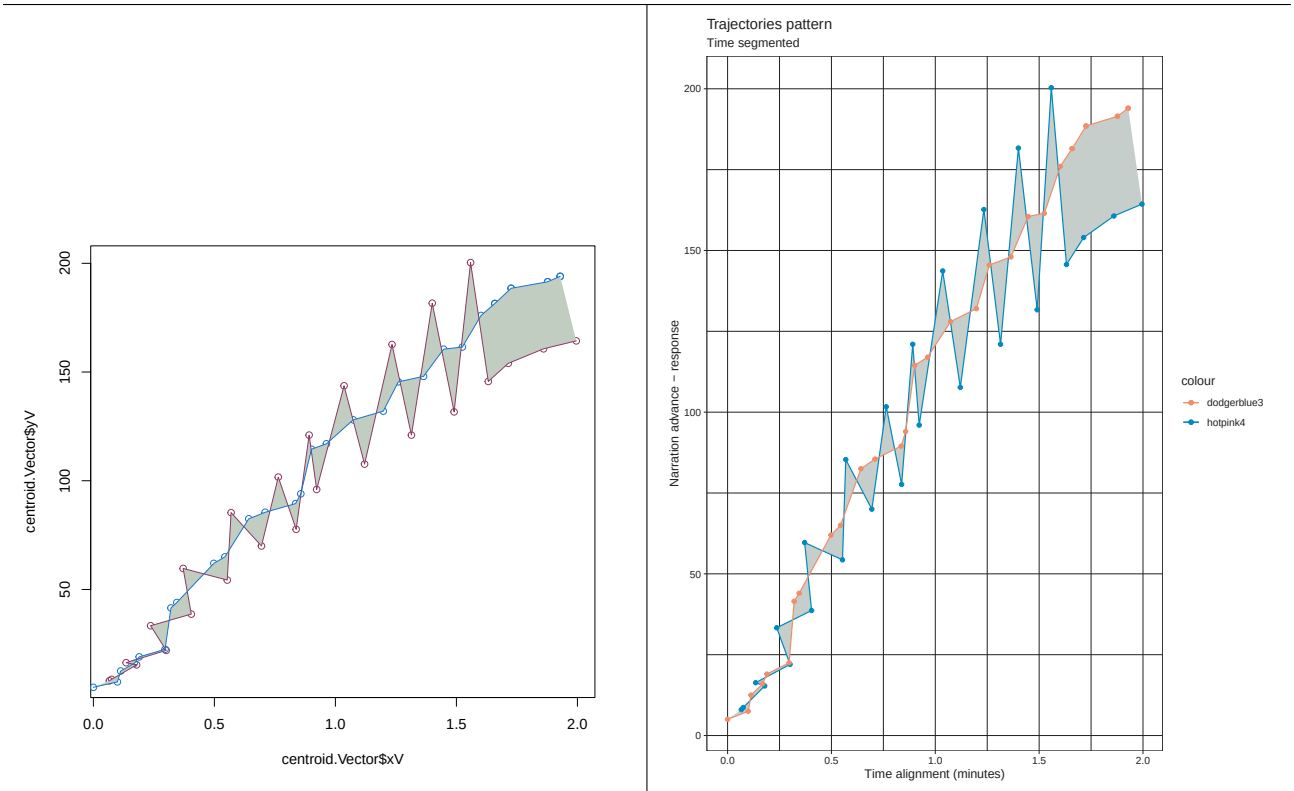


(c), Triangle-segmented



(d), Patterns

Figure 5.7. Trajectories - proximity



(a), plot-polygon() **(b), ggplot-geom_polygon**

Figure 5.8. Trajectories - centroids vs midpoints curves proximity

is *smoother* than that of the center of mass, as shown in the figure 5.7(d). However, the midpoint curve could also *hide* slight adjustments in the weighted language.

So, to integrate these two new referent trajectories, *we consider the area between the path generated by the centroids and the path derived from the midpoints as our proximity measure*, as shown in figure 5.8. The basis is that this bi-dimensional measure contains an appropriate notion of the proximity between the pair of original trajectories. With this, instead of evaluating the proximity at discrete points, we are observing the complete development of the trajectories.

Measure Computation. On the one hand, we calculate the area bounded by two curves based on its definition, i.e. the integral of the absolute value of the difference between the two trajectories (equation 1)

$$f(x) = \int_{v_1}^{v_2} |f_2 - f_1| dx \tag{1}$$

For this, we need to interpolate the trajectories to have the values of both in the same points of the shared interval, taken as 25 equally-spaced steps. The number of steps was set conventionally. Then the integral (eq.1) of the difference of those values is calculated.

We compared the calculated value by the definition against an implementation of a function in a library. This required normalized data. So, we divide the values of X and Y by the maximum of each set to have them in the unit interval. Contrasting the results, we noticed that the employed package does not consider the absolute value in the computation. The values calculated using the concept of integration but removing the absolute value are very similar to these. Particularly, for the variable *total_words*, there was a difference of 0.00011098 between the two calculated values that remains below the determined absolute error (< 0.00012).

5.3.3 Basis of comparison.

We found two measures in the literature, *Hausdorff* and *Fréchet* (section 2.3.2), suggested to estimate the similarity between two trajectories. They are based on the distance that the curves maintain in their progress. Since they are defined in terms of distance, we select them to estimate proximity in our context. So, the next step is to compare the calculated values in section 5.3.2 with the proposed evaluations of Hausdorff and Fréchet, and examine the variation of the correlation among features, to analyze the variability in their relationship.

Correlation tracking To track the correlation, we consider the scales of *Pearson, Spearman*, and *Kendall*. They showed to be consistent for different types of linguistic features and configurations, when they were employed in the characterization of the alterations in the language after a TBI [58]. With their respective considerations, depending on the features behavior, for example, the Pearson coefficient ($\wp$), will usually show linear associations, while Spearman(ρ) and Kendall(τ) are for a broader type of association [59].

5.4 Articles & Conferences

1. Roldán-Palacios M., López-López A. "Understanding Syntax Structure of Language After a Head Injury", Brain Informatics (Lecture Notes in Artificial Intelligence), vol. 13406, pp.288–300

2. An invited poster derived from the article "Understanding Syntax Structure of Language After a Head Injury" was submitted to Peeref (https://www.peeref.com/). Published on April 28, 2023 (DOI: https://doi.org/10.54985/peeref.2304p7848418).

6 Preliminar discussion

Although experimentally we have not covered the complete methodology proposed, we have advanced in it.

1. The definition of characteristics, planned to be complete in the academic term we ended, is extended due to recent updates available on TalkBank. This involves reviewing it and adjusting our feature set if necessary.

2. The process we have defined for preprocessing, time integration, subsample generation, and feature extraction must be applied for the remaining instances per participant for the different samples of the recovery period and the different feature types.

3. In the identification of the comparison measures, we must review and solve the correct segmentation in triangles for the scenarios, beyond the ideal case that allowed an immediate partition of the area between the compared trajectories, alternating their vertices.

4. Regarding the comparison bases that we established, Hausdorff and Fréchet, we found that despite the first one is defined in terms of topological spaces and the second one in the optimization of a shortest path, when applying them to our context, we get the same value. Therefore, the comparisons will be completed in reference to Hausdorff. Besides, the consistency between Hausdorff and Fréchet will be reviewed, since there is a precedent in another context in which both measures maintain the same value.

5. At the moment we have worked on the experiments with correlations between features identified in previous works, now it is time to define clustering criteria since the comparison measures have been identified.

6. The representations and values to be evaluated will be generated as the different stages of the experimentation are completed.

7. With a scheme defined in the experiments carried out, we must complete enough to identify the interval of variability as well as the value ϵ for its evaluation.

8. We have completed a first general statement of the sought technique, which in parallel to the progress in each worked stage must be adjusted as required.

6.1 Further results & discussion

6.1.1 Experimental Setting

In our subsequent experimental setting (figure 6.1), we worked on study and control samples with 5 participants for each case. We completed the comparison between the features of one of the defined lexical sets, comprising 'Total.Words' versus 'Duration..sec.', 'Total.Utts', 'Words.Min', 'MLU.Words', after applying the steps described above, we calculated the measurements to be compared.

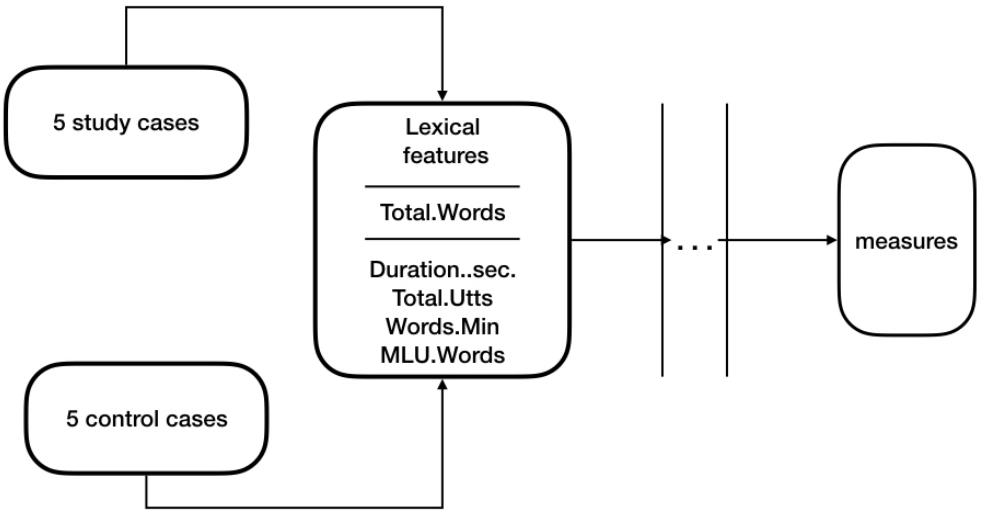


Figure 6.1. Next experimental setting.

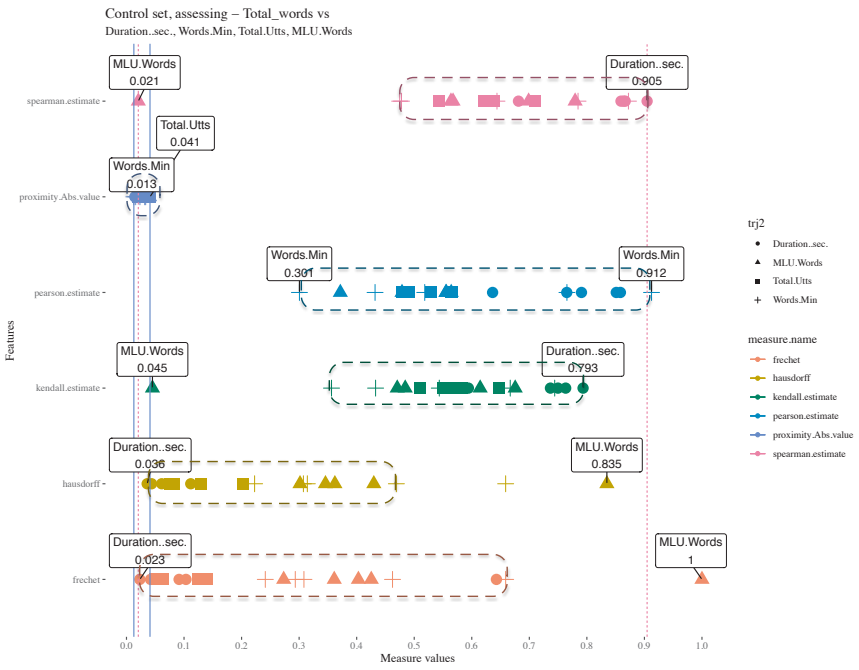
6.1.2 Results

Figures 6.2(a) and 6.2(b) summarize the obtained results from these calculations, which correspond to control and study groups, respectively. Where X-axis consists of the calculated values, the Y-axis corresponds to the responses by measures. Besides, the shape of the point indicates the lexical feature.

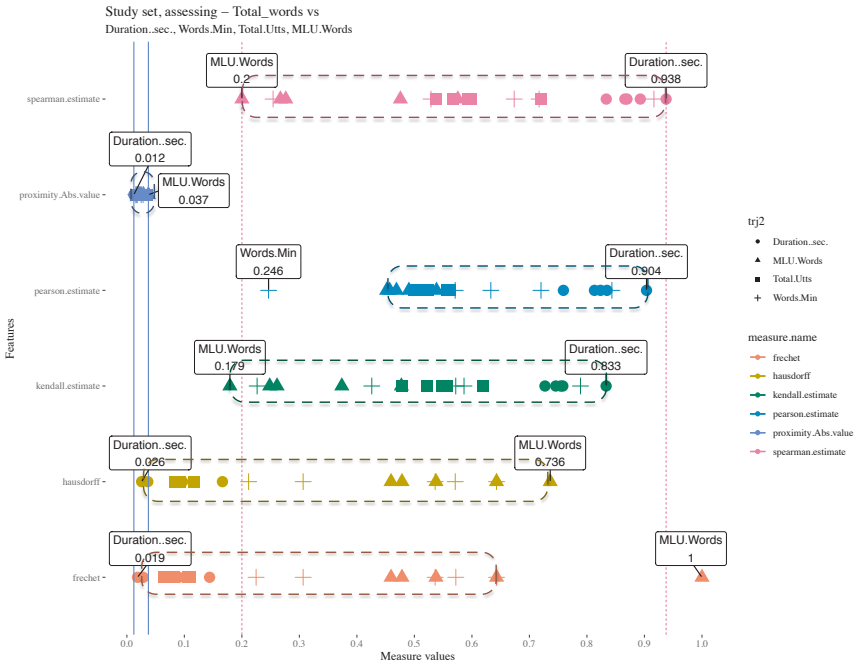
6.1.3 Discussion

About the control cases, a gap between the minimum and the rest of the responses for Spearman's evaluation is notable, which makes us consider an outlier, and the same is observed for Kendall's evaluation. Furthermore, Hausdorff and Fréchet behave this way to their maximum values, consistent with their definition, given that we seek minimum values for them. Also, note that this outlier occurs for the same feature, 'MLU.Words'. Pearson maintains a more or less uniform distribution where it varies. On another side, let us observe that the proposed proximity measure restricts to a short range.

For the study group, the condensed response of the proposed measure replicates here, although in a smaller range. We highlighted the intervals in which their values are mostly condensed for each group to compare information for the study and control cases. We have notable changes in the distribution of the responses of the characteristics comparing both groups. For the study case, Pearson and the proposed measure of proximity vary in smaller limits, contrary to Spearman, Kendall, and Hausdorff, where the intervals for the control cases are smaller than those for the study cases. Fréchet, although it presents changes in the response of the characteristics, mainly for 'MLU.Words' and 'Words.Min', practically remains in the same range of variation. Figure 6.3 help us to extend the discussion to the proximity measure. Once again, the shape of the point indicates the characteristic against which the total number of words compares. The graphic representation



(a), Measures results for Control set



(b), Measures results for Study set

Figure 6.2. Measure comparison - Control & Study cases

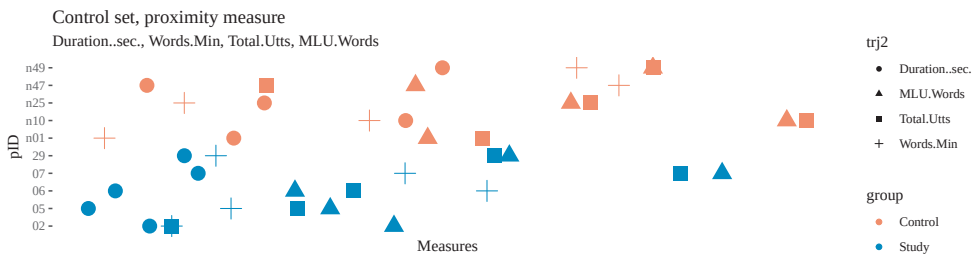


Figure 6.3. Proximity measure

grouped by participant allows us to observe that the responses that deviate from the general behavior of the attributes belong to the same participant for both study and control cases. Besides, the evaluations for 'Duration..sec.' are concentrated to the far left for the study set. Furthermore, for the control cases, there is a more uniform distribution of the responses of the attributes throughout the interval in which it varies. These last results lead us to additional open questions, such as *(i)* What does a greater distance between the centroids and the midpoints mean? or *(i)* Do control participants have more defined profiles (narrative style)?

References

[1] D. K. Menon, K. Schwab, D. W. Wright, and A. I. Maas, "Position statement: Definition of traumatic brain injury," *Archives of Physical Medicine and Rehabilitation*, vol. 91, no. 11, pp. 1637–1640, 2010.

[2] K. Blennow, D. L. Brody, P. M. Kochanek, H. Levin, A. McKee, G. M. Ribbers, K. Yaffe, and H. Zetterberg, "Traumatic brain injuries," *Nature Reviews Disease Primers*, vol. 2, no. 16084, 2016.

[3] A. Brazinova, V. Rehorcikova, M. S. Taylor, V. Buckova, M. Majdan, M. Psota, W. Peeters, V. Feigin, A. Theadom, L. Holkovic, and S. Anneliese, "Epidemiology of traumatic brain injury in europe: A living systematic review," *Journal of Neurotrauma*, vol. 38, pp. 1411–1440, 2021.

[4] M. C. Dewan, A. Rattani, S. Gupta, R. E. Baticulon, Y.-C. Hung, M. Punchak, A. Agrawal, A. O. Adeleye, M. G. Shrime, J. V. Rubiano, Andrés M. Rosenfeld, and K. B. Park, "Estimating the global incidence of traumatic brain injury," *Neurosurgery*, vol. 130, pp. 1080–1097, 2019.

[5] C. Iaccarino, A. Carretta, F. Nicolosi, and C. Morselli, "Epidemiology of severe traumatic brain injury," *Neurosurgical Sciences*, vol. 62, no. 5, pp. 535–541, 2018.

[6] S. Ahmed, H. Venigalla, H. Madhuri Mekala, S. Dar, M. Hassan, and S. Ayub, "Traumatic brain injury and neuropsychiatric complications," *Indian J Psychol Med*, vol. 39, pp. 114–121, 2017.

[7] A. R. Rabinowitz Ph.D. and H. S. Levin Ph., "Cognitive sequelae of traumatic brain injury," *Psychiatr Clin North Am.*, vol. 37, no. 1, pp. 1–11, 2014.

[8] E. Power, S. Weir, J. Richardson, D. Fromm, M. Forbes, B. MacWhinney, and L. Togher, "Patterns of narrative dis- course in early recovery following severe traumatic brain injury," *Brain Injury*, vol. 34, no. 1, pp. 98–109, 2020.

[9] B. MacWhinney, *The CHILDES Project, Tools for Analyzing Talk*. Mahwah, NJ., Lawrence Erlbaum Associates, 3rd ed., 2000.

[10] E. Stubbs, L. Togher, B. Kenny, D. Fromm, M. Forbes, B. MacWhinney, S. McDonald, R. Tate, L. Turkstra, and E. Power, "Procedural discourse performance in adults with severe traumatic brain injury at 3 and 6 months post injury," *Brain Injury*, vol. 32, no. 2, pp. 167–181, 2018.

[11] M. Roldán-Palacios and A. López-López, "A. feature analysis for aphasic or abnormal language caused by injury.," *Intelligent Computing. Lecture Notes in Networks and Systems.*, vol. 285, pp. 1–16, 2021.

[12] H. Dhrif and S. Wuchty, "Stable feature selection for gene expression using enhanced binary particle swarm optimization," *Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020)*, vol. 2, pp. 437–444, 2020.

[13] P. Laube, *Computational Movement Analysis*. SpringerBriefs in Computer Science, Cham: Springer International Publishing, 2014.

[14] A. E. Ramage, "Potential for cognitive communication impairment in covid-19 survivors: a call to action for speech-language pathologists," *American Journal of Speech-Language Pathology*, vol. 29, no. 4, pp. 1821–1832, 2020.

[15] L. Cummings, "Pragmatic impairment and covid-19," *Intercultural Pragmatics*, vol. 19, no. 3, pp. 271–297, 2022.

[16] K. Chadd, K. Moyse, and P. Enderby, "Impact of covid-19 on the speech and language therapy profession and their patients.," *Frontiers in Neurology*, vol. 12, no. 629190, pp. 1–11, 2021.

[17] A. Köse, H. Tayyip Uysal, M. Merve Parlak, A. Baştug Dumbak, M. Tanrıverdi, and K. Mariam, "The investigation of the cognitive communication functions of survivors of coronavirus disease 2019 (covid-19): A survey study," *Karya Journal of Health Science*, vol. 3, no. 3, pp. 338–342, 2022.

[18] L. Al-Shalabi, "New feature selection algorithm based on feature stability and correlation," *IEEE Access*, vol. 10, pp. 4699–4713, 2022.

[19] U. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review.," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, 2022.

[20] L. Bryant, A. Ferguson, and E. Spencer, "Linguistic analysis of discourse in aphasia: A review of the literature.," *Clinical Linguistics & Phonetics.*, 2016.

[21] L. Bryant, E. Spencer, and A. Ferguson, "Clinical use of linguistic discourse analysis for the assessment of language in aphasia," *Aphasiology*, vol. 31, no. 10, pp. 1105–1126, 2017.

[22] M. J. Benacquista and J. D. Romano. Undergraduate Lecture Notes in Physic, Springer, 2018.

[23] A. Piper, R. J. So, and D. Bamman, "Narrative theory for computational narrative understanding," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, vol. 18, no. 1, pp. 298–311, 2021.

[24] F. Pianzola, "Looking at narrative as a complex system: the proteus principle," *Narrating complexity*, pp. 101–122, 2018.

[25] G. A. DiLisi, *Classical Mechanics, Volume 1, Tools and vectors*. IOP Concise Physics, 1210 Fifth Avenue, Suite 250, San Rafael, CA, 94901, USA: Morgan & Claypool, 2019.

[26] M. v. Kreveld, M. Löffler, and L. Wiratma, *On Optimal Polyline Simplification using the Hausdorff and Fréchet Distance*, vol. 99 of *SoCG 2018*, pp. 56:2–56:14. LIPIcs, June 2018.

[27] H. Su, S. Liu, B. Zheng, X. Zhou, and K. Zheng, "A survey of trajectory distance measures and performance evaluation.," *The VLDB Journal*, vol. 29, no. 1, pp. 3–32, 2020.

[28] M. Pertti, vol. 44 of *Cambridge studies in advance mathematics*. Cambridge University Press, 1995.

[29] Y. Liang, W. Chen, and W. Cai. Fractional Calculus in Applied Sciences and Engineering, De Gruyter, 2019.

[30] H. Alt and M. Godau, "Computing fréchet distance between two polygonal curves," *International Journal of Computational Geometry & Applications*, vol. 5, no. 1–2, pp. 75–91, 1995.

[31] S. Wang, Z. Bao, J. Culpepper, and G. Cong, "A survey on trajectory data management, analytics, and learning.," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–36.

[32] E. Belogay, C. Cabrelli, U. Molter, and R. Shonkwiler, "Calculating the hausdorff distance between curves," *Information Processing Letters*, vol. 64, no. 1, 1997.

[33] B. Aronov, S. Har-Peled, C. Knauer, Y. Wang, and C. Wenk, "Fréchet distance for curves, revisited," *arXiv preprint arXiv:1504.07685v1.*, 2018.

[34] J. Kloke and J. W. McKean, *Nonparametric Statistical Methods Using R*. Chapman & Hall/CRC The R Series, CRC Press, 2015.

[35] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces. knowledge and information systems," *Knowledge and information systems*, vol. 12, pp. 95–116, 2007.

[36] W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald, and N. Amri, "A review of the stability of feature selection techniques for bioinformatics data," IEEE *13th International Conference on Information Reuse & Integration (*IRI*)*, pp. 356–363, 2012.

[37] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review.," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, 2022.

[38] W. Al-Saeedan and M. E. B. Menai, "Swarm intelligence for natural language processing," *International Journal of Artificial Intelligence and Soft Computing*, vol. 5, no. 2, pp. 117–150, 2015.

[39] L. M. Abualigah, A. T. Khader, and E. S. Hanandeh, "A hybrid strategy for krill herd algorithm with harmony search algorithm to improve the data clustering," *Intelligent Decision Technologies*, pp. 1–12, 2017.

[40] L. Wiratma, *Computations and Measures of Collective Movement Patterns Based on Trajectory Data*. PhD thesis, Universiteit Utrecht, December 2019.

[41] L. Wiratma, M. v. Kreveld, and M. Löffler, *On Measures for Groups of trajectories*, pp. 311–330. Lecture Notes in Geoinformation and Cartography, Springer, 2017.

[42] K. H. Almotairi and L. Abualigah, "Hybrid reptile search algorithm and remora optimization algorithm for optimization tasks and data clustering," *symmetry*, vol. 14, no. 458, pp. 1–29, 2022.

[43] C. A. Coelho, B. Grela, M. Corso, A. Gamble, and R. Feinn, "Microlinguistic deficits in the narrative discourse of adults with traumatic brain injury.," *Brain Injury*, vol. 19, no. 13, pp. 1139–1145, 2005.

[44] L. Togher, E. Elbourn, B. Kenny, E. Power, S. McDonald, R. Tate, L. Turkstra, F. D. Holland, A., M. Forbes, and B. MacWhinney, "Tbi bank is a feasible assessment protocol to evaluate the cognitive communication skills of people with severe tbi during the subacute stage of recovery.," *Brain Injury*, vol. 28, no. 5-6, pp. 723–723, 2014.

[45] J. Steel and L. Togher, "Social communication assessment after traumatic brain injury: A narrative review of innovations in pragmatic and discourse assessment methods.," *Brain Injury*, vol. 33, no. 1, pp. 48–61, 2019.

[46] J. Steel, E. Elbourn, and L. Togher, "Narrative discourse intervention after traumatic brain injury: A systematic review of the literature.," *Topics in Language Disorders*, vol. 41, no. 1, pp. 47–72, 2021.

[47] L. E. Hanna, "The relationship between discourse- and sentence-level processing during narrative production following traumatic brain injury," Master's thesis, 2018.

[48] M. Pond, "Story grammar analysis of cinderella narratives in adults with and without traumatic brain injury," Master's thesis, Spring 2020.

[49] E. Elbourn, B. Kenny, E. Power, C. Honan, S. McDonald, R. Tate, A. Holland, B. MacWhinney, and T. Leanne, "Discourse recovery after severe traumatic brain injury: Exploring the first year. brain injury," *Brain Injury*, vol. 33, no. 2, pp. 143–159, 2019.

[50] T. Niu and M. Bansal, "Adversarial over-sensitivity and over-stability strategies for dialogue models," *arXiv preprint arXiv:1809.02079*, 2018.

[51] J. Chen, D. Tam, C. Raffel, M. Bansal, and D. Yang, "An empirical survey of data augmentation for limited data learning in NLP," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 191–211, 2023.

[52] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners. advances in neural information processing systems," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[53] M. A. Hedderich, L. Lange, A. Heike., J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," *arXiv preprint arXiv:2010.12309*, 2020.

[54] O. Melamud, M. Bornea, and K. Barker, *Combining unsupervised pre-training and annotator rationales to improve low-shot text classification*, pp. 3884–3893. November.

[55] I. Burak Ozyurt, "On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining," *bioRxiv*, pp. 2020–05, 2020.

[56] K. Lë, C. Coelho, and J. Fiszdon, "Systematic review of discourse and social communication interventions in traumatic brain injury," *American Journal of Speech-Language Pathology*, vol. 31, no. 2, pp. 99–1022, 2022.

[57] B. MacWhinney, *Tools for Analyzing Talk - Electronic Edition Part 2: The CLAN Programs*. Carnegie Mellon University, electronic ed., 2020.

[58] M. Roldán-Palacios, "A multi-level analysis of language production: Features compromised by traumatic brain injury.," Master's thesis, INAOE, 2021.

[59] M.-T. Puth, M. Neuhäuser, and G. D. Ruxton, "Effective use of spearman's and kendall's correlation coefficients for association between two measured traits," *Animal Behaviour*, vol. 102, pp. 77–84, 2015.