Disfluency as an Indicator of Cognitive-Communication Disorder Through Learning Methods

Marisol Roldán-Palacios and Aurelio López-López(✉)

Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro No. 1, Sta. María Tonantzintla, 72840 Puebla, Mexico
{marppalacios,allopez}@inaoep.mx
https://ccc.inaoep.mx/

Abstract. The analysis of the different varieties of language alterations from several causes has become an indicator to support tentative diagnoses, not only physical but degenerative, functional or cognitive. In this study, we explore fluency-disfluency in language of participants after suffering a traumatic brain injury. From a linguistic-computational approach, covering one-year of periodic post-recovery stages samples, candidate subsets of features were evaluated with a pool of learning methods until obtaining comparable scores to a baseline taken as the maximums achieved with the same evaluation, but on the full feature set. Starting in three-months recovery stage, this was extended to six, nine, and twelve months. After setting a global overview during this period of the fluency response based on F1-score of the learning algorithms, the identified feature was the basis to work on a model in a longitudinal sense of the disfluency-response with dichotomous global linear mixed effects model.

Keywords: Cognitive-communication disorder · Traumatic brain injury · Disfluency · Machine learning · Global linear mixed effect models

1 Introduction

"Language is one of the most important products of human cerebral action, but also because the problems raised by the organization of language seem to me to be characteristic of almost all other cerebral activity." Lashley-1951 [9].

Cognitive-communication disorder is the term coined to describe anomalies in language as a consequence of a traumatic brain injury (TBI) [1,14]. Although *Western Aphasia Battery-Revised* is the most accepted instrument to examine a part of the affectations on language [2] caused by TBI, there are some sequelae not revealed by that test [17]. Besides, fluency has been identified as a possible factor guiding the exploration of the atypical language following a TBI [14].

Supported by CONACyT and partially by SNI.

M. Mahmud et al. (Eds.): BI 2021, LNAI 12960, pp. 46–56, 2021.
https://doi.org/10.1007/978-3-030-86993-9_5

Regarding fluency alteration, this work aims to find a feature subset of the indices measuring fluency, in addition to reveal changes across periodic stages of recovery, sampled after the injury. To achieve this, a principal component analysis along with the corresponding correlation analysis were performed to find the features contributing the best information for the discrimination task, and sensitive enough to reflect subtle modifications.

An evaluation step of the selected subset employing varied learning algorithms was added to determine the final selected subset. The results reported in this work consider three algorithms: Random Forest, SPAARC and Naïve Bayes. The approach followed to achieve a longitudinal response predictor was to elaborate a model based on the features evaluated with the learning methods considering multilevel modeling [7].

The contributions of our work are the following: a) The identification of a reduced group of *four features* of language fluency showing an F1 score above the whole feature set; b) we show that a learning method model based on trees operates accordingly to the worked context, language variations, in a wrapping step to evaluate this sort of variables; and c) A fused approach to define a part of the TBI-language reactions, first with learning methods to identify indicative features. Then, feeding them to a mixed model to know how language re-adapt during the recovery stage, a period barely studied.

The organization of the paper is as follows. After summarizing related work in Sect. 2, the initial whole feature group is described in Sect. 3. Then, experiments Sect. 4 includes the data description, its pre-processing, the methodology, selected feature set, results, discussions about features and learning methods, and ending with a revision of the longitudinal model applied. The work closes with conclusions and considerations of further analyses in Sect. 5.

2 Related Work

There is no consensus about the reliability of the inspection of features like repetitions, revisions and fillers (mazes) [17] in language impairment analysis. These and few other associated with disfluency were disregarded when assessing language discerning alterations in some cognitive skill after TBI [8], or revising discourse performance based on meaningful words [17].

While considering mazes, regression models were built to approximate an understanding of cognitive impairments related to sentence planning deficits observed in TBI-language [12]. Additionally, the number of fillers and abandoned words as speech fluency barometers, along other features, were examined with methods based on learning and language models, comparing their efficacy to determine language impairment [6].

After learning models were introduced [13], fed by language measures, to discern among different neuro-degenerative conditions, this angle has continued growing [3–5,15]. Such approach can complement those studies in which, declarative and working memory, attention, executive functions and social condition as *cognitive constructs* have been substantially investigated, and that started to

shed some insights in their relationship with (non simulated) functional use of language [16] in TBI cases.

3 Feature Group

The package *flucalc* [10,19] extracts a pool of more than forty attributes. Among them, we find *#TD*, i.e. typical disfluencies (by definition the sum of phrase repetitions, word revision, phrase revision, pause counts, and filled pauses), *%TD* corresponding to the total typical disfluencies over the total words or total syllables, *#SLD* described as stutter-like disfluencies, including the sum of prolongations, broken words, words, part-word repetition (PWR), phonological fragments, and monosyllabic whole word repetition (WWR) along with *%SLD* proportion *#SLD* in reference to the total intended words. The *SLD Ratio* is calculated as $SLD/(SLD + TD)$, and the measure of *weighted SLD*, which is a relatively complex relation involving additions, subtractions, products and a ratio of PWR, mono-WWR, PWR-RU, mono-WWR-RU, prolongations and blocks, where *RU* stands for *repetition units*. All these features in addition to the measures in which they are based on, and some more, are addressed to assess fluency in the altered language. The complete list and description are in [11].

4 Experiments

The information collected for the experiments carried out for the analysis is reported with more detail in the next subsections, starting by the data set.

4.1 Data Collection

Regarding the studied group, a detailed description of the project of the data corpus [17] is given in [14]. Briefly, this consists of samples elicited to a selected cohort group of participants after being affected by a TBI, registered at three, six, nine, and twelve months, after the injury. The group consists of few more than fifty participants, however, a missing stage of recovery sample was allowed due to exceptional circumstances, that leaves an average of forty participants per period, mostly male.

From several tasks, the *recount of Cinderella story* was selected for this analysis. The negative set came from a different investigation [1], where the *generative story based on a picture* task was taken. Both data sets are in TBIBank [11,18]. So both task samples corresponding to study and negative cases respectively, were transcribed from recorded speech instances. Fluency attributes set, as described above, are obtained from those transcripts worked by experts.

4.2 Pre-processing

The indices extracted from *flucalc* package [11,19] condense densities, ratios, additions or other composed functions as *weighted SLD* (described in Sect. 3). These were pre-processed with a simple transformation to leave them all in a comparable interval, where the set of features varies.

4.3 Methodology

As shown in Fig. 1, the methodology starts with a *principal components analysis* (PCA) on the first recovery phase (i.e. the three months sample), the next step was to determine how many components derived from PCA were those contributing most, according to the problem at hand. Analyzing the evolution of language after the brain injury, during post-traumatic stages, we determined to keep as much information as possible. Based on customary *elbow graph* (Fig. 3) of a total of forty, the first twelve components were chosen after which the trajectory relatively stabilizes to a constant, covering so far 0.944 of the variance computed by PCA. Then, we established which features of each of the twelve linear combinations selected were strongest related to their corresponding PC and were taken as the initial subset. Noisy or neutral features were removed by correlation analysis. By the same criteria, the subset was extended, to then ablate while they are evaluated with learning methods to set the definitive selection.

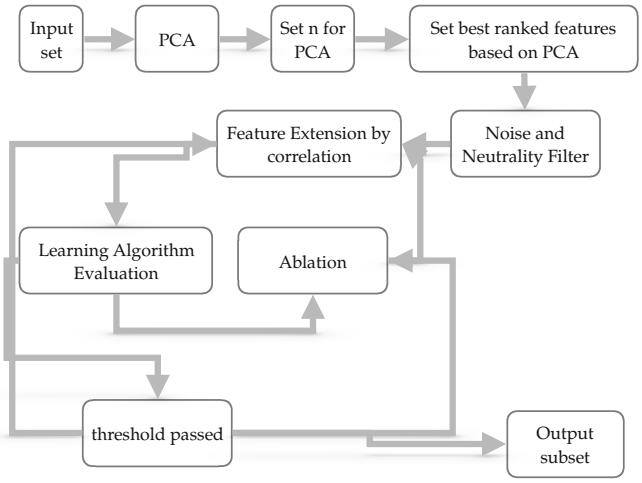

Fig. 1. Flucalc - feature selection - methodology

4.4 Selected Feature Set

Explicitly, the initial subset is $S_1 = \{mor\_Utts, mor\_Words, tot\_Prolongation, tot\_WWR, tot\_Phrase\_repetitions, prop\_Word\_revisions, prop\_Pauses, prop\_TD, Content\_words\_ratio\}$. The *prop_Broken_word* attribute, though associated to *PC6* was not included in that subset due to a poor correlation exhibited between them, along with the fact that it stays in the negative pole with respect to PC12. Promptly, *Content_words_ratio* and *tot_Prolongation* attributes were

removed having a noisy or neutral influence with the rest. After shrinking and expanding successively the feature group, the definitive set is detailed in Table 1.

Table 1. Flucalc 4 features subset

FN	PC	Feature	Description
1	PC12	mor_Utts	Total utterances in the sample
24		Mean_RU	= (PWR-RU + WWR-RU)/(PWR + WWR)
27	PC4	tot_Phrase_repetitions	Total phrase repetitions
30	PC8, PC9	prop_Word_revisions	Proportion of word revisions

[a]RU, repetition units
[b]PWR, part-word repetition
[c]WWR, whole word repetition

Observe that *mor_Utts*, one of the most contributing feature, was selected after noting that was related to *PC12*. Having an individual weight of over 75 of efficacy, *Mean_RU* attribute was added because of a missing correlation with those in the revised subset. This consists of the ratio between the sum of the *part-word repetition units* added to the index of the *whole-word repetition units* and the addition of *part-word repetition* and the *whole-word repetition* (Table 1).

4.5 Results

The response of the 4-features subset describing fluency in comparison with full-set performance on the inspected instances are summarized in Fig. 2. This consists of the curve of evaluation of the selected features along side the base-curve calculated for the full feature set, both evaluated with *Näive Bayes, SPAARC*, and *Random Forest* methods, for the corresponding three, six, nine and twelve months of recovery. This gives a look of language changes after TBI.

An average of forty instances per group per period were evaluated with a *Leave-One-Out Cross Validation* (LOOCV) scheme. Periods of recovery are expressed in the X-axis and macro *F1-score* are plotted in Y-axis.

4.6 Feature Selection Discussion

We start by observing the principal components (PCs) related to three of the determined feature subset, showed in second column of Table 1. There, we can notice that they are distributed along the whole group of PCs considered, i.e. they do not lead the PC list. For instance, the total whole word repetitions *tot_WWR* associated with PC1 was removed due to having a negative role in the subset, when evaluated with the learning algorithms. Counter-intuitively, we did not find that some subset of measures defining typical disfluencies could lead the discriminating task. The first correlation showed certain negative effect of those features on some PCs then, learning algorithms evaluations evidence that densities and relations or functions associated with the concept of *typical disfluencies*

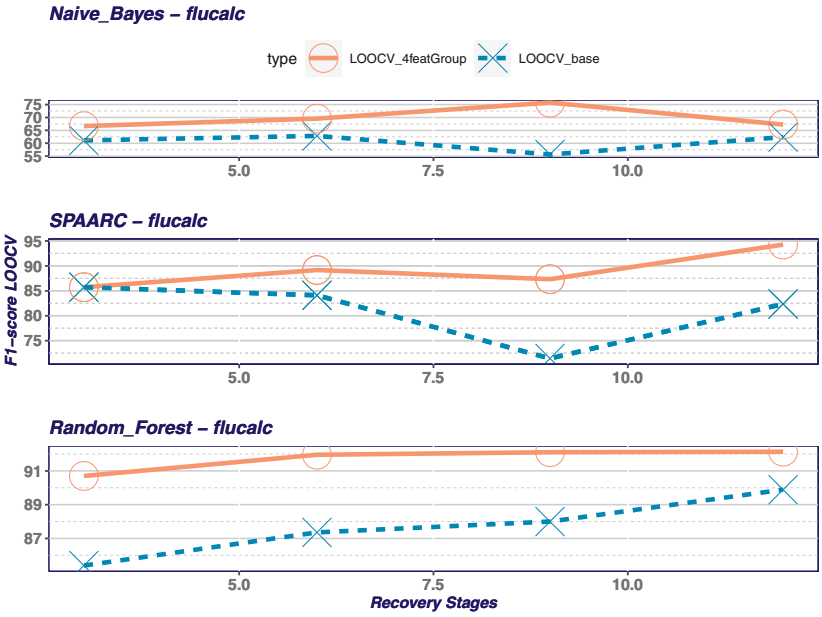

Fig. 2. (a) Naïve Bayes - (b) SPAARC - (c) Random Forest - fluency response - first year after traumatic brain injury

(TD) [11], in addition to those based on the concept of *stutter-like disfluencies* (SLD) do not work well together. In other words, they add noise if evaluated as a set along with other features, though individually some of them are relatively contributors for the task at hand. Further steps in the process led us to *Mean_RU* that was not revealed by any linear combination from the PCA analysis. Figure 2 shows that: $\forall p_i, F1_{score}(\text{4-features subset}) > F1_{score}(\text{full features})$, where $p_i \in \{recovery\_periods\}$, i.e. for each stage of recovery evaluated. This indicates that the $F1$-score curve corresponding to the 4-feature subset remains above the $F1$-score of the full-feature group, for each of the methods illustrated, suggesting that the 4-features subset encompasses the whole information of the full feature *flucalc* set. For instance, *Random Forest* is moving in the $[90.70, 92.13]$ interval regulating the discrimination task of negative versus study cases, and remains still sensitive to reflect the subtle changes in the *cognitive-communication disorder* analyzed with grounds on the fluency of the TBI-language samples inspected. Observing that none of the *TD* or the *stuttered-like disfluencies* [11], neither any composed function related to them, are part of this *4-features subset* (Table 1), together with the fact that none of them is correlated to PC1 or PC2 from PCA, provide evidence of the complexity in the characterization [14] of the *reorganized* language following a TBI.

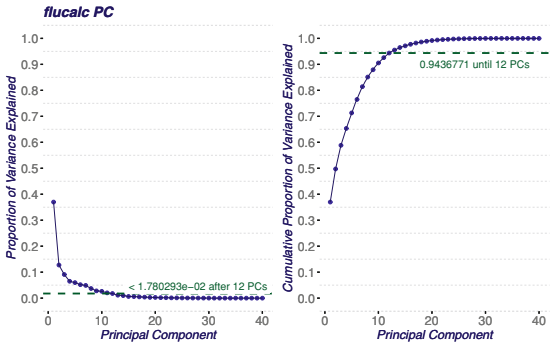

Fig. 3. 4-features subset - flucalc - correlation at three-months recovery stage.

4.7 Learning Algorithm Evaluation Discussion

Seven learning algorithms were taken into account for the evaluation step of the features: Sequential Minimal Optimization, Naïve Bayes, Bayes Net, SPAARC, Random Forest, Classification via Regression, and Adaboost. From these, results for the first, third and fourth are reported here. To judge how much information the *flucalc* feature set contains, the set was evaluated as a whole, and taken as baseline. A tenth (Table 1) of the original feature set [11] allows to achieve comparable efficacy measures in contrast with those obtained from the entire *flucalc* feature set. Though the comparison is illustrated with $F1$-score, calculated as $F1\text{-score} = \frac{2*TP}{2*TP+FP+FN}$, where TP, FP, FN respectively mean True-Positives, False-Positives and False-Negatives in the *confusion matrix*, to appropriately reflect the proportion of the elements in this latter. The *F1-score* values range in the $[0, 1]$ interval, but here they are expressed as percentages.

Contrasting response curves, *Random Forest* is indicating a more consistent behavior for both samples, negative versus study cases for the complete group of time points. The *base-set* curves move in a relatively similar scale for both tree based learning algorithms, but *SPAARC* reveals the existence of noisy features with a fall of around 15% in $F1$-score for the nine months stage of recovery. This is also evident in the trajectory that the response follows for the baseline set evaluated with *Naïve Bayes*, which in fact relatively mirrors the reaction illustrated for the flucalc base set evaluated with *SPAARC*, though the former moves to lower values than the latter. A common behaviour exhibited by the three learning methods is that *4-features subset* curve remains above the baseline trajectory for the four time points samples of the first year following TBI.

From what was described above, we can state that the discriminating task can be done with the chosen *4-features subset*. Moreover, though *Naïve Bayes* algorithm generally replicates the behavior between *4-features subset* and *baseline* curves, i.e. there is a gap between them, the former rests on the latter for the whole period considered, *baseline* moves in an interval around 60%, not giving much certainty about the registered efficacy. *SPAARC* and *Random Forest* keep consistency for both trajectories not only in the described aspects but both get

acceptable measures, however, their response for each recovery stage oscillates dis-similarly, which does not allow to suggest anything regarding an amelioration or decline in the *cognitive-communication disorder* caused by a TBI.

4.8 Pattern Complexity

As a contribution to outline the intrinsic complexity of language in a *cognitive-communication disorder*, we include Fig. 4. Where every curve represents individual response by stage of recovery at 3, 6, 9, and 12 months. Each row represents one of the four features selected {*mor_Utts, Mean_RU, tot_Phrase_repetitions, prop_Word_revisions*}. The left side depicts TBI-language instances and right side draws negative responses. Same participants were evaluated in each side, though some graphs appears to have less, this is caused by the pattern of response. Presenting a more varied density, *mor_Utts* and *prop_Word_revisions* features allow to distinguish more than one response pattern. In the former, at least two clusters, one group of responses relatively stable remaining under 0.2 index and the other turning up and down from one time point to the next. The latter registers at least three patterns, one relatively steady resting below 0.25, one alternating valleys and peaks, and one more growing in the first three recovery stages and then drops. In *tot_Phrase_repetitions* attribute graph, in a minor or major grade. An S response group is differentiated from the cluster resting on zero. *Mean_RU* seems to exhibit two more conducts in addition to that leaving on zero, one for above 0.3 estimation and the other oscillating between zero and values equal or greater than this estimate. One last finding is that study group and negative cases seems to be comparable in density by feature.

4.9 Longitudinal Model

The next step was to work in determining the simplified expression to predict language disfluency-fluency over time. For that purpose, a longitudinal random intercept model with dichotomous response was tried, the elementary with only two nested structures, the repeated appraisals and the time. The features fed to the suggested model were the evaluated with the learning methods previously described, the four selected features *4-features set* = {*mor_Utts, Mean_RU, tot_Phrase_repetitions, prop_Word_revisions*}, in addition to the time points for the fixed effect part with random intercept in terms of a generalized linear mixed effects model. As proof of significance was applied, a *likelihood ratio test* which can be explained as the comparison of the likelihood of two models. Both inputs raise from the same structure but one *with* and the other *without* the factors of interest, the latter named *the null model*. The difference between these two models determines if a *fixed effect/variable* becomes significant, if the former is significant the latter will be. From one side, from the manageable tests, any confirmation of significance of the inspected characteristics was obtained, from another, examination brought to a singularity problem. The absence of some samples per period per participant in the current sample was one of the obstacles, given that the present study works in language impairment observing for

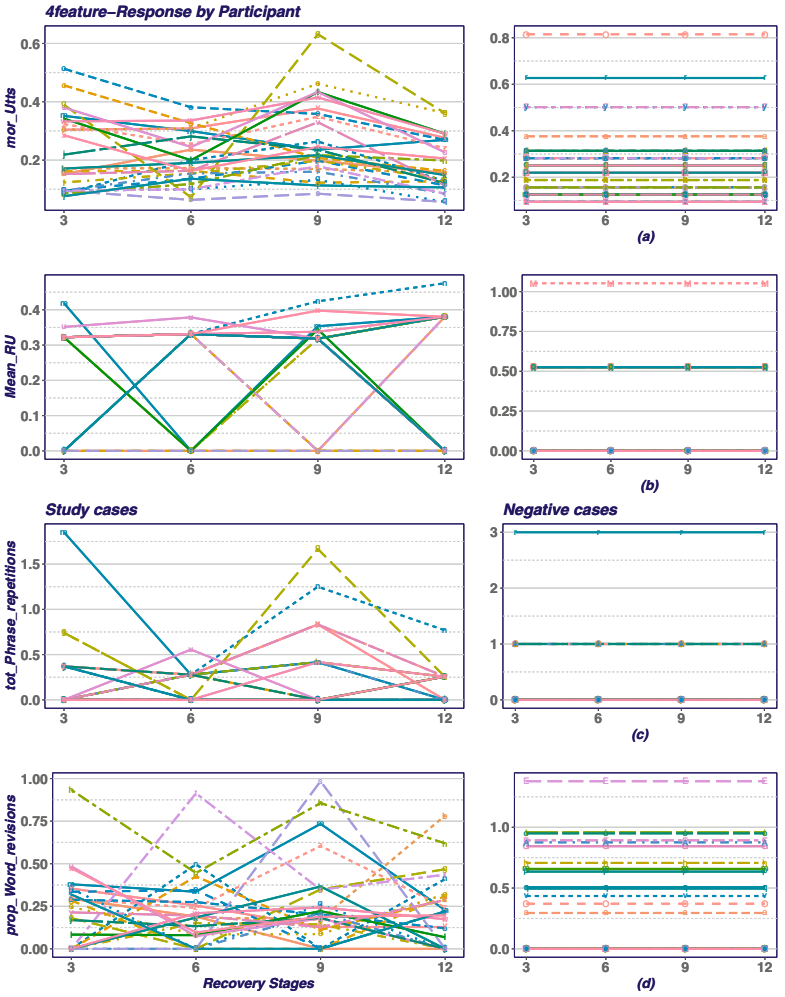

Fig. 4. (a) mor_Utts - (b) Mean_RU - (c) tot_Phrase_repetitions - (d) prop_Word_revisions - individual fluency response by feature - first year after TBI

any subtle adjustment in it, this could not be suitable to try any technique of data augmentation due to the inherently bias added and removing those incomplete records left different conditions for the learning algorithms in reference to multilevel modeling, situating us right beyond any determination of the complementarity of both techniques.

5 Conclusions and Further Work

The assessment in terms of $F1$-score supports the selection of the 4-features subset with indices moving in the interval $[90.70, 92.13]$ for Random Forest and

above 85 for SPAARC learning methods. Furthermore, the subset showed to include the information provided by the whole fluency feature.

However, setting a direct determination of a longitudinal model implementing a multilevel approach to predict the response across time was not completed, in part due to the combined techniques handle different conditions on the fed data. Results obtained removing noisy and neutral characteristics suggests that this is a proper approach to recognized contributors features, the extended analysis to predict TBI-language response over time have to be solved with a trade-off between a more scarcely data and the reliability of the results based on them.

A limitation of the followed approach is that, though results were summarized in F1–score, an additional detailed assessment of the learning methods can be carried out. Additionally, the sensitive factors of learning methods and mixed models have to be considered in advance to allow a less intricate *flow* of data from the learning algorithms to mixed effects model, to reach a good approximation of the response of the studied TBI-affected language.

References

1. Coelho, C.A., Grela, B., Corso, M., Gamble, A., Feinn, R.: Microlinguistic deficits in the narrative discourse of adults with traumatic brain injury. Brain Inj. **19**(13), 1139–1145 (2005)
2. Elbourn, E., et al.: Discourse recovery after severe traumatic brain injury: exploring the first year. Brain Inj. **33**(2), 143–159 (2019)
3. Fraser, K.C., et al.: Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. Cortex **55**, 43–60 (2014)
4. Fraser, K.C., Hirst, G., Graham, N.L., Meltzer, J.A., Black, S.E., Rochon, E.: Comparison of different feature sets for identification of variants in progressive aphasia. In: Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Baltimore, MD, USA, 27 June 2014, pp. 17–26 (2014)
5. Fraser, K.C., Hirst, G., Meltzer, J.A., Mack, J.E., Thompson, C.K.: Using statistical parsing to detect agrammatic aphasia. In: Proceedings on Biomedical Natural Language Processing (BioNLP), Baltimore, MD, USA, pp. 134–142 (2014)
6. Gabani, K., Sherman, M., Solorio, T., Liu, Y., Bedore, L.M., Peña, E.D.: A corpus-based approach for the prediction of language impairment in monolingual English and Spanish-English bilingual children. In: The 2009 Annual Conference of the North American Chapter of the ACL (NAACL), Boulder, CO, pp. 46–55 (2009)
7. Hair, J.F., Jr., Fávero, L.P.: Multilevel modeling for longitudinal data: concepts and applications. RAUSP Manag. J. **54**(4), 459–489 (2019)
8. Jorgensen, M., Togher, L.: Narrative after traumatic brain injury: a comparison of monologic and jointly-produced discourse. Brain Inj. **23**(9), 727–740 (2009)
9. Lashley, K.S.: The problem of serial order in behavior. In: Jeffress, L.A. (ed.) Cerebral Mechanism in Behavior, pp. 112–136. Wiley, New York (1951)
10. MacWhinney, B.: The Childes Project: Tools for Analyzing Talk, 3rd edn. Lawrence Erlbaum Associates, Mahwah (2000)
11. MacWhinney, B.: Tools for analyzing talk - electronic edition part 2: the CLAN programs. Carnegie Mellon University (2020)

12. Peach, R.K.: The cognitive basis for sentence planning difficulties in discourse after traumatic brain injury. Am. J. Speech Lang. Pathol. **22**, S285–S297 (2013)
13. Peintner, B., Jarrold, W., Vergyri, D., Richey, C., Gorno-Tempini, M.L., Ogar, J.: Learning diagnostic models using speech and language measures. In: 30th Annual International IEEE EMBS Conference, Vancouver, BC, Canada, pp. 20–24 (2008)
14. Power, E., et al.: Patterns of narrative discourse in early recovery following severe Traumatic Brain Injury. Brain Inj. **34**(1), 98–109 (2020)
15. Rentoumi, V., et al.: Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: a computational linguistics analysis. In: 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), pp. 11–14 (2017)
16. Rowley, D.A., Rogish, M., Alexander, T., Riggs, K.J.: Cognitive correlates of pragmatic language comprehension in adult traumatic brain injury: a systematic review and meta-analyses. Brain Inj. **31**(12), 1564–1574 (2017)
17. Stubbs, E., et al.: Procedural discourse performance in adults with severe traumatic brain injury at 3 and 6 months post injury. Brain Inj. **32**(2), 167–181 (2018)
18. TBI bank. https://tbi.talkbank.org/. Accessed 3 Mar 2021
19. Childes Project. https://talkbank.org/. Accessed 3 Mar 2021