Charles University

Faculty of Arts

Doctoral dissertation

Mgr. Michal Láznička

Discourse Production of Czech Speakers with Aphasia

A Usage-based Exploration

Produkce diskurzu českých mluvčích s afázií: Explorace s využitím usage-based lingvistiky

Department of Linguistics

Study program: Filologie (obecná lingvistika)

Dissertation supervisor: Mgr. Eva Lehečková, Ph.D.

2022

I hereby declare that I have written this dissertation independently, using only the mentioned and duly cited sources and literature, and that the work has not been used in another university study programme or to obtain the same or another academic title.

In Prague on September 21 2022 Mgr. Michal Láznička

Abstract

The research in linguistic aphasiology has been dominated by structuralist, rule-based approaches to the study of langauge. However, recent work has shown that analyses based in constructivist, usage-based frameworks can provide explanations to patterns of language processing in aphasia that are difficult to accommodate in structuralist models. The present work follows up on these findings and aims to provide additional evidence for the benefits of the usage-based model by using data from Czech speakers with aphasia, an understudied language in this context. The aims of the study were threefold: to create a collection of samples of aphasic connected speech available to other researchers, to provide a description of the patterns of aphasic discourse production in Czech, and, most importantly, to show potential benefits of usage-based construction grammar for aphasia research.

A corpus of the speech of eleven persons with fluent and non-fluent aphasia of varying degrees of severity was created. The corpus consist of more than 23000 word position produced by speakers with aphasia in tasks used to elicit conversational, narrative, descriptive, and procedural discourse. The corpus is lemmatized and morphologically tagged and the transcripts are aligned with audio recordings. A smaller sample of three, demographically matched neurotypical speakers is also included. A sample of the corpus with a more detailed annotation was used in subsequent analyses.

First, a quantitative description of the micro-structure of Czech aphasic discourse production was carried out. A suite of measures of fluency, productivity, and well-formedness differentiated between a group of persons with non-fluent aphasia on the one hand and a group of neurotypical speakers and individuals with fluent aphasia on the other. Consistent with the current state of knowledge, non-fluent speakers produced shorter utterances with fewer verbs and a high number of disfluencies and sentence fragments. A hierarchical clustering analysis that revealed several language profiles within the fluent and non-fluent group.

The second analytical part consists of three case studies that are concerned with the role of cumulative lemma frequency, relative frequency of a paradigmatic cell, and frequency of cooccurrence in aphasic speech production. An analysis of verb production found that lemma frequency can better explain observed differences on group and individual level, compared to structural complexity. Participants with more severe word finding problems relied more on high frequency verbs and non-fluent participants produced verbs with complex argument structure frames that simultaneously have high frequency. They also relied more on the use of the existential-presentative construction which is both highly frequent and partially lexically fixed. Verbs with higher frequency also induced fewer disfluencies.

An analysis of inflected nouns in the corpus showed that speakers with more severe aphasia produced fewer inflected word forms which was caused by their frequent use of single word utterances or the existential-presentative construction. Nevertheless, these speakers were able to spontaneously produce inflected forms that had high relative frequency or were tied to specific argument structure constructions.

A third analysis of the production of prepositional phrases found that combinations of particular prepositions and nouns that are frequently used together were produced more fluently which points to the fact that such multiword units are stored as chunks and retrieved as read-made wholes rather than assembled from component parts during production. This was true even in phrases with prenominal modifiers that were overall produced with more disfluencies.

These finding demonstrate the benefits of the usage-based framework for the analysis of patterns of langauge behavior in aphasia and have implications for clinical practice that could benefit from integrating frequency effects in the formulation of assessment tools and approaches to therapy.

Keywords:

acquired aphasia, chunking, corpus linguistics, discourse production, disfluencies, fluency, frequency effects, inflectional morphology, usage-based construction grammar

Abstrakt

Výzkum v lingvistické afaziologii využíval po dlouhou dobu především strukturalistické přístupy založené na pravidlech. Některé výsledky z poslední doby však ukazují, že konstruktivistické přístupy založené na užívání jazyka (usage-based přístup) dokážou vysvětlit některá specifika zpracování jazyka v afázii, která jsou ve strukturalistickém rámci obtížně vysvětlitelná. Předkládaná dizertační práce navazuje na tyto výzkumy a klade si za cíl předložit další důkazy pro výhodnost usage-přístupu. Využívá přitom data z češtiny, která je v afaziologickém výzkumu značně podreprezentovaná. Práce si stanovila tři cíle: jednak shromáždit projevy českých mluvčích s afázií, které by byly přístupné dalším výzkumníkům, dále podat detailní popis produkce diskurzu v afázii v češtině a konečně ukázat některé přednosti usage-based přístupu pro afaziologii.

V rámci práce byl vytvořen korpus jedenácti mluvčích s fluentní a nefluentní afázií s různými stupni závažnosti poruchy. Korpus obsahuje přes 23000 slovních pozic vyprodukovaných mluvčími s afázií sebranými s využitím úkolů, jejichž cílem bylo elicitovat konverzační, narativní, deskriptivní a procedurální diskurz. Korpus je lematizován a morfologicky označkován. Dále je v něm zahrnut menší vzorek řečové produkce tří neurotypických mluvčích se srovnatelnými demografickými charakteristikami. Část korpusu byla opatřena detailnější anotací a využita v následných analýzách.

V první analytické části byl proveden kvantitativní popis produkce diskurzu českých mluvčích s afázií. Soubor vybraných proměnných popisujících fluentnost, produktivitu a gramatičnost dokázal rozlišit mezi mluvčími s nefluentní afázií na jedné straně a skupinou neurotypických participantů a mluvčích s fluentní afázií. Získané jazykové profily jsou srovnatelné s podobnými výzkumy z jiných jazyků. Nefluentní mluvčí produkovali kratší výpovědi s nižší frekvencí sloves a vyšším množstvím dysfluencí a větných fragmentů. Následná klastrová analýza odhalila v rámci skupin specifické jazykové profily.

Druhá analytická část sestává ze tří případových studií, které jsou zaměřeny na roli frekvence lemmatu, relativní frekvence paradigmatické buňky v rámci lemmatu a kookurenční frekvence v jazykové produkci mluvčích s afázií. Analýza produkce lexikálních sloves ukázala, že kumulativní frekvence lemmatu vysvětluje pozorované rozdíly mezi jednotlivými participanty lépe než strukturní komplexnost. Participanti s těžší poruchou lexikálního zpracování používali více sloves s vysokou frekvencí. Nefluentní mluvčí produkovali slovesa s komplexnější valenční strukturou, která však byla zároveň i více frekventovaná. Tito participanti také více využívali existenciálně-prezentativní konstrukci, která je vysoce frekventovaná a její lexikální obsazení je zároveň z části pevně dané. Ukázalo se také, že slovesa s vyšší frekvencí byla produkována s méně dysfluencemi.

Analýza skloňovaných tvarů substantiv v korpusu ukázala, že mluvčí s těžšími projevy afázie produkovali méně nenominativních tvarů, což souvisí s tím, že používali více jednoslovných výpovědí a existenciálně-prezentativních konstrukcí. Tito mluvčí nicméně byli schopni spontánní produkce vyskloňovaných tvarů, které mají vysokou relativní frekvenci slovního tvaru nebo se objevují ve specifických valenčních konstrukcích.

Třetí analýza se týkala produkce předložkových frázích. Výsledky ukazály, že konkrétní kombinace předložek a substantivních doplnění, které mají vysokou frekvenci užívání jsou mluvčími s afázií

produkovány s menším množstvím dysfluencí. Toto zjištění platí i pro předložkové fráze s prenominálními modifikátory, které byly všeobecně produkovány méně fluentně. Tyto výsledky naznačují, že podobné vysoce frekventované víceslovné jednotky jsou v jazykové znalosti uloženy jako chunky a jsou vybavovány jako celek spíše než generovány z jednotlivých částí.

Předkládané výsledky ukazují výhody usage-based přístupu pro analýzu specifik jazykového chování v afázii. Zároveň mají důsledky pro klinickou praxi, která by mohla s výhodou zařadit důsledky frekvenčních efektů v jazyce při konstrukci diagnostických nástrojů a plánování terapie.

Klíčová slova:

chunking, dysfluence, flektivní morfologie, fluentnost, frekvenční efekty, korpusová lingvistika, produkce diskurzu, usage-based konstrukční gramatika, získaná afázie

Table of Contents

Index of Tables

Table of Figures

1 Introduction

Aphasia is a condition that has been closely related to the study of the neurobiology of language ever since the pioneering works of Wernicke, Steinthal, or Broca. Later on it has emerged as one of the sources of evidence to support some of the theoretical constructs and processes hypothesized by theoretical linguists. However, first and foremost, it is a condition that severely impacts the quality of life of the persons suffering from it as well as their significant others. It is primarily for these reasons that it deserves the attention of linguists. Not only can linguistics improve our understanding of language and its functioning by applying its models and their predictions to aphasic data, but, more importantly, the descriptive and analytical apparatus of linguistics can contribute to the development of more sensitive clinical tools and, consequently, to the improvement in quality of life of the persons who have aphasia.

Linguistic aphasiology has formed as a field of study during the 1950s and 1960s and has been dominated by structuralist, "rules-and-lexicon" approaches to language. However, studies employing cognitive-functionalist, constructivist, and usage-based perspectives have started to emerge recently that show promising results, indicating that some of the assumptions of these approaches may uncover patterns in the linguistic behavior of persons with aphasia that would not be of interest to structuralist frameworks. Similarly, the descriptive and analytical concepts and tools of these approaches offer straightforward explanations to some of the findings concerning language processing in aphasia that are difficult to accommodate in structuralist frameworks.

While the interest in the linguistic analysis of aphasia has been on the rise ever since the 1960s, the bulk of research has focused on English in particular and a small sample of the major languages of Western Europe in general. Czech is particularly under-researched in this regard and the lack of linguistic descriptions and analyses has hindered the development of clinical tools that would account for the structural-typological peculiarities of Czech. The lack of research on semi-spontaneous connected speech of Czech speakers with aphasia is particularly striking.

The present work aims to address these two facts and open a new avenue of research that sets the course to fill this much needed gap. In order to achieve these goals, I first present a corpus of aphasic speech which I assembled to provide other researchers with a data set to use for their analyses, to generate hypotheses, and to contribute to with own data. Using this data, I proceed to provide a first quantitative characteristic of aphasic discourse production in Czech. Finally, I explore the way in which usage-based construction grammar, a framework grounded in functionalist, cognitive linguistic assumptions, can be used to analyze aphasic data and serve as a means to generate conclusions with important implications for clinical practice. The work is structured as follows.

In the first part, I provide a definition of aphasia and a basic historical overview of linguistic research of aphasia (sections 2.1 and 2.2). Particular attention is paid to studies that have used aphasic data to provide arguments for a particular theory of language representation and processing or that have employed such theories to explain the processing patterns observed in different types of aphasia. This section is concluded with a remark on the importance of cross-linguistic research of aphasia. I then proceed with a general discussion of available studies of aphasia in Czech (section 2.3). The general introduction proceeds with an overview of the basic assumptions and tenets of usage-

based construction gramar and, finally, with a discussion of the body of usage-based research on aphasia (sections 2.4 and 2.5). I conclude the general introduction with a practical demonstration of the way in which aphasic data can be approached using the usage-based framework.

The second part describes the construction of the corpus of aphasic discourse production and its basic characteristics (section 3). This description is followed by an analysis of discourse production of eleven persons wih aphasia and three neurotypical speakers using a suite of variables measuring fluency, productivity, and well-formedness (section 4.2). This characteristic is followed by three case studies aimed to show the benefits of integrating the usage-based model, particularly different measures of frequency as the variables central to this model, in the analysis of aphasic data. Sections 4.3.1 and 4.3.2 describe the production of verbs and inflected nouns in the corpus with special focus on lemma and word form frequency. The last case study (section 4.4) is concerned with the analysis of disfluency production in the context of prepositional phrase production, particularly in relation to cooccurrence frequency. The analysis of verb and noun production as well as the patterns of disfluency in prepositional phrases suggest that not only does cumulative lemma frequency play a role in lexical retrieval in aphasia, but that relative frequency of individual inflectional variants as well as the probability of cooccurrence of two word forms modulates the ease of production and fluency in aphasia. These findings are further discussed in the context of language assessment and therapy in concluding remarks.

The present work is meant both for readers with background in usage-based linguistics and construction grammar as well as researchers of aphasia and specialists from clinical practice. This has impacted the way the text is structured as I have attempted to keep the work accessible to these diverse groups of prospective readers. The result of this approach is that some of the sections may be a touch too introductory for the respective groups. Prospective readers are kindly invited to skim through or skip such passages altogether.

2 General Introduction

eventualni miniodstavec kde se popise pojeti a struktura cely tyhle kapitoly

2.1 Aphasia: definition, syndromes, and symptoms

In this section, I am going to very briefly define aphasia and provide an overview of aphasia types as they will be presented in this work. Then I move on to the relationship between aphasia and linguistics, providing arguments for a closer cooperation between the two fields. In particular, I focus on the discussion of agrammatic aphasia and its relevance to general models of language representation and processing.

A standard definition of acquired aphasia which is the topic of the present work states that it is an acquired neurogenic language specific disorder (Hallowell & Chapey 2008). This means that the language capacity is affected selectively as a cognitive domain and general intellect and other cognitive capacities are spared. Aphasia is caused by lesions to the brain regions that are involved in language processing, i.e., in particular, the so called perisylvian regions in the frontal and temporal lobes of the left hemisphere. This implies that the consequent language impairment is not caused by sensorimotor problems. Aphasia is most frequently caused by stroke (cerebrovascular accident). The prevalence of stroke in Czechia is between 200-300 patients per 100,000 persons (Kalita et al. 2013; Sedova et al. 2017), the number of people with aphasia in Czechia is, to the best of my knowledge, not available.[1] However, it is estimated that aphasia is present in 21-38 % of acute stroke patients (Berthier 2005). Code and Petheram (2011) estimate that the incidence of aphasia in the developed world is between 0.02-0.06 % with prevalence of 0.1-0.4 %. These numbers make it clear that this condition, with its direct negative impact on quality of life (cf. e.g. Carota et al. 2001), is a topic of high importance. Other causes of aphasia include traumatic brain injuries, tumors, or inflammations, aphasic symptoms may also occur in neurodegenerative diseases. However, most studies in aphasia research only include stroke-induced aphasia, as this patient group is much larger than the others and post-stroke aphasia is typically caused by well-defined focal lesions (Hallowell & Chapey 2008) that provides greater methodological validity to group comparisons.

Aphasia may impact all levels of linguistic structure, language production and comprehension, and both spoken or signed as well as written language.[2] Table 1 is an overview of typical language symptoms that occur in aphasia. Czech examples are provided from the corpus assembled as one of the outcomes of the present work.

1 The authors of AfaSlovník (Macková, Macko & Hanibalová 2020), a web-based therapy tool, claim that approximately 5000 individual affected by aphasia every year. However, no source for this figure is provided.

2 Given the incidence estimates for aphasia and the numbers of sign language users, the population of signers with aphasia is understandably small. See Hickok and Bellugi (2001) for an overview of aphasia in sign languages.

symptom	characteristic	example
anomia	lexical retrieval problem, slowed or failed retrieval marked by silent and filled pauses	*a on na něho H H H on 1 se rozběhl* 'and he toward him H H H he 1 started running' (ba2: 14; target = *rozběhl se* 'start running')[3]
circumlocution	when lexical retrieval fails, the target word is described by means of its formal and/or functional characteristics	*začala H musela spát* 'she began H had to sleep' (aa3: 88; target = *omdlít* 'faint'
agrammatism	omission of function words and/or inflectional morphemes that renders an utterance ungrammatical	*bych si n- 0.5 najít* '(I) would f- 0.5 find' (aa3: 159; target = *bych si musel najít* 'would need to find')
paragrammatism	substitution of function words and/or inflectional morphemes that renders an utterance ungrammatical	*kočka se všímla že ...* 'the cat REFL-ACC noticed that ...' (pa1: 82; target = *všimla si* 'notice REFL-DAT'
phonemic paraphasia	lexical retrieval error resulting in the production of formally similar word or nonword	*venku 0.2 H byl 1 pech pes* 'outside 0.2 H there was 1 a dock dog' (aa3: 187, target = *pes* 'dog')
semantic paraphasia	lexical retrieval error resulting in the production of semantically similar word	*pes ne ale 0.2 kráva ne* 'dog no but 0.2 cow no' (aa4: 9; target = *osel* 'donkey')
neologistic paraphasia	lexical retrieval error resulting in the production of a formally unrelated nonword	*má 0.2 má 0.5 H 3 peram* 'has 0.2 has 0.5 H 3 *peram*' (aa2: 249; target unknown)
conduite d'approche	a gradual production of words or word fragments that are formally similar to the target word; typical symptom of conduction aphasia	*myč moš myč 0.2 myche 0.5 myš* (aa3: 283; taget = *myš* 'mouse')

Table 1: Typical aphasic symptoms with examples from the data; participant id, c-unit id, and target word are in brackets.

Aphasia can be classified into different clinical profiles or subtypes based on the localization of the lesion and the dominant symptoms. From the anatomical point of view aphasia can be categorized based on the localization of lesions in the frontal (typically inferior frontal gyrus) or temporal (typically superior temporal gyrus) lobe. The most coarse-grained, behaviorally based division is between fluent and non-fluent aphasia. Fluent aphasia is roughly characterized by word finding difficulties, frequent paraphasias, paragrammatic errors and comprehension problems, whereas non-fluent aphasia is defined by reduced mean length of utterance, frequent silent and filled pauses, and agrammatic errors. Historically, a dichotomy between motor and sensory aphasia was also used which is based on impaired language production and comprehension respectively. Roughly speaking, frontal lesions correlate with non-fluent and motor syndromes, while temporal lesions are con-

3 The information in brackets such as ba2: 14 indicates the id of the participant and the number of utterance in which the example was produced. The examples were translated to correspond to the original as closely as possible. Morphological segmentation and glossing is only provided when essential for the argument presented by the example. All the examples from the corpus presented in the text use the following notation:
- H = hesitation sound
- 0.2 = pause longer than 0.2 seconds; the following thresholds (in seconds) are used to differentiate between pauses are 0.5, 1, 3, 5, 10
- n- = word fragment
- <cough> = all other non-hesitation paralinguistic sounds are transcribed in this way

nected to fluent and sensory types. Despite the fact that aphasia is characterized by a great amount of variation, both intra- and interindividually,[4] there are certain clinical profiles or prototypical clusters of symptoms that frequently cooccur and that correlate with lesion sites (Hallowell & Chapey 2008; Damasio 2008). This has led to more detailed classification systems of which the Boston classification has been the most influential. The Boston classification system of aphasia types was developed on the basis of frequent lesion-symptom mappings by the Boston group, a group of aphasiologists and neurologists around DF Benson, N Geschwind, H Goodglass, E Kaplan, and others (e.g. Benson & Geschwind 1971; or Goodglass & Kaplan 1972a). The system, in simplified terms, classifies patients based on the state of the following language components: fluency, comprehension, repetition, and naming. Table 2 shows how the combination of these variables defines the subtypes. The second influential and widely used classification is Luria's system (Luria 1964; Valdois, Ryalls & Lecours 1989; Akhutina 2016) which has been used in Eastern Europe, Russia, or Latin America.

aphasia type	fluency	comprehension	repetition	naming
Anomic	+	+	+	−
Wernicke	+	−	−	−
Conduction	+	+	−	−
Sensory transcortical	+	−	+	−
Broca	−	+	−	−
Motor transcortical	−	+	+	−
Mixed transcortical	−	−	+	−
Global	−	−	−	−

Table 2: Aphasia types with four characteristic features; + stands for spared and - for impaired ability.

2.2 Aphasia and linguistic theory

Language is the cognitive domain affected by aphasia and it is logical that the relationship between linguistics and aphasia should be a tight one. However, this has not been, and to an extent is not, always the case. Before I proceed with a brief history of this relationship from the theoretical linguistic perspective in the second half of the 20th century, I would like to explicitly state where the mutual benefits lie for both disciplines. From both a clinician's and an aphasia researcher's perspective,

4 This has led some authors to reject finer-grained classification in favor of the coarse-grained types mentioned which are then to be combined with detailed individual profiles (Caramazza & McCloskey 1988; Marshall 2010). For instance, Ardila (2010) claims that there are only two clearly identifiable syndromes, a Wernicke-like impairment of paradigmatic processing and a Broca-like impairment of syntagmatic processing in the sense of Jakobson's theory of aphasia (1995).

there is a need for at least a rudimentary descriptive apparatus that is needed to describe a person with aphasia's linguistic behavior. Such description can be theory-neutral, or theoretically shallow and only make use of a set of basic notions that would be based on a particular descriptive tradition of the language. To take Czech as example, this would be a basic inventory of word classes and inflectional categories that can be found in secondary education grammar books. This is mostly the case in the context of speech language therapy in Czechia.[5]

Alternatively, one can adopt a specific theory or model of language representation and processing. Such an approach has, potentially, two merits. Firstly, a model of language may serve as a basis for generating predictions and explanations of expected and observed linguistic behavior in aphasia. For instance, by adopting the usage-based model and factoring in constructional frequency, one can explain why some lexically specific instantiations of the passive construction may be comprehended with relative ease despite the fact that this structure is notoriously difficult for people with aphasia (cf., among many others, e.g. Caramazza & Zurif 1976; or Cho & Thompson 2010).[6] Such a model also provides predictions with regards to the structures that can be expected to be resistant or susceptible to errors in aphasic language processing. For instance, when one looks at the production of Czech prepositional phrases in aphasia, which demands the retrieval of a particular preposition and an appropriate inflected word form of the complement noun, combinations of particular prepositions and nouns that have a high probability of cooccurrence would be predicted to be less prone to error (cf. section 4.4.2 of the present work).

An application of a specific model of language in therapy has the potential to provide more sensitive and reliable assessment and therapy tools. When testing performance on specific structures, such as passives or inflected forms under the usage-based model, it makes sense not only to test if a patient performs differently from neurotypical individuals on the structure as a whole, but to also factor in different levels of difficulty based on usage factors such as frequency. Under the assumption that the differences between persons with aphasia and neurotypical individuals are quantitative rather than qualitative in nature (e.g. Gahl & Menn 2016, see also section 2.5 below), it is expected that instantiations of a construction that are easier to process for neurotypical speakers will also be less difficult for people with aphasia. I believe it has clinical value to assess the severity of impairment for individual patients based on their performance on a scale of difficulty. The same principle may be applied in the design of therapeutic tools. While working within a very different framework from the one adopted in the present work, Springer and colleagues (2000), Thompson and Shapiro (2005), or Kiran et al. (2012) demonstrate the rationale and potential merits of employing a linguistically informed approach to therapy.

From the perspective of theoretical linguistics, the use of aphasic data also has clear benefits, primarily in model building and "criticism". If we assume that one of the aims of linguistic research is to develop a model of language representation and processing, such models should not only account for data of neurotypical, adult native speakers, but also for other groups, such as children in the process of L1 acquisition, non-native speakers, or speakers with language impairments such as aphasia. A general model of language representation and processing should be able to explain why

5 This information is based on my personal communication with several speech language therapists during the course of the project presented here.

6 See also the discussion in sections 2.4-2.6.

certain phenomena are more prone to error in aphasia or to predict potential impairment of a language component based on an observation of a different impaired component that is according to a given model subserved by the same mechanism. Aphasia has played such a role in some areas of psycholinguistic research where patterns of errors are used to argue for specific components of models of language processing (e.g. Dell et al.'s model of lexical retrieval 1997) and, most prominently, in neurolinguistic research where mappings between lesion sites and symptoms serve as evidence to support models of the neurobiology of language.[7] However, it is still relatively rare for general psycholinguistics and even more so for theoretical linguistics to reflect on impaired data. The remainder of this section is dedicated to a brief discussion of research that used aphasic data to argue for a specific model of language representation and processing.

Descriptions of syndromes with the manifestations typical for aphasia and reflections of the relationship between language impairments and brain lesions appear throughout history since the ancient Egyptian, Greek, and Roman period (Benton & Joynt 1960; Benton & Anderson 1998; Code 2013; Tesak & Code). The foundations of the modern scientific study of aphasia were laid in the 19th and early 20th century with works of Wernicke, Broca, Lichtheim, and others. However, the descriptions and theorizing of these early researchers contain only crude and pre-theoretical concepts with regards to language. This is not surprising, given that these were mostly neurologists and neuropsychiatrists and the fact that the linguists of the time were mostly concerned with historical and comparative aspects of language. A Pick was one of the first researchers of aphasia to include a model of language and its relation to aphasia in his work (1913).

However, it was not until Jakobson that theoretical linguistics found interest in aphasia and linguistic aphasiology became one of the major branches of aphasia sciences in the second half of the 20th century. Roughly speaking, Jakobson's model of aphasia (1964; 1980; 1995) was based on two assumptions. First, he argued that the order of acquisition of linguistic categories is negatively correlated with their disruption in aphasia such that early acquired phenomena are less impaired in aphasia. He also defined two aphasic syndromes based on two basic relations between components of language, viz. paradigmatic and syntagmatic relations. Impairment of the paradigmatic axis results in selection or substitution errors, whereas impaired syntagmatic relations block the ability to combine linguistic items and thus produce coherent, grammatical utterances. These two syndromes roughly overlap with Wernicke's and Broca's aphasia respectively.

It was Broca's aphasia that had become of major interest to linguists concerned with aphasia in the 1970s and 80s. The then prevailing understanding was that lesions to the frontal regions of the brain result in motor aphasias characterized by effortful production with reduced length of utterance and frequent omission of function words, but spared comprehension, while temporal lesions cause sensory aphasias with impaired comprehension and relatively spared production with paraphasias and substitution errors. Caramazza and Zurif (1976) published a paper that initiated a shift in this view and elevated the interest of linguists in aphasia. Caramazza and Zurif tested the comprehension of three groups of individuals with Broca's, Wernicke's, and conduction aphasia on syntactically com-

7 In fact, post mortem investigations of lesioned brain structure combined with recorded observational data of a given individual's linguistic behavior were, prior to the advance of neuroimaging methods, the only source of model building in neurobiology of language (e.g. Wernicke 1974).

plex sentences with center embedded relatives that were semantically reversible, irreversible, or had an “improbable” interpretation (see Table 3).

condition	example stimulus
semantically reversible	The cat that the dog is biting is black.
semantically irreversible	The apple that the boy is eating is red.
improbable	The dog that the man is biting is black.
control	The cat is chasing a blue bird.

Table 3: Conditions and example stimulus sentences from Caramazza and Zurif (1976)'s experiment.

They found that the performance of the participants with Broca’s aphasia was comparable to neurotypical speakers on control and irreversible sentences, but dropped on reversible and improbable sentences where it compared to the performance of the group with Wernicke’s aphasia. Their conclusion was that contrary to previous claims, comprehension in Broca’s aphasia is not entirely spared and that in contexts that require a syntactic analysis for correct comprehension, such as semantically reversible structures, their performance is similar to language production in that these individuals not only produce agrammatic or *asyntactic* structures but also show a pattern of *asyntactic* comprehension.

This conclusion was of particular interest to the generative theory, the dominant paradigm at the time, because it could be interpreted as a double dissociation between syntax and lexical semantics. Broca’s aphasia, at the time believed to be caused mainly by lesions to Broca’s area, appeared to be a syndrome characterized by a general impairment of the grammatical competence systems with agrammatic but semantically intact production (e.g. *Dog chase cat* for a target *The dog chases the cat*) and asyntactic comprehension (e.g. *The cat is chased by the dog* interpreted as *cat chase dog*). Conversely, Wernicke’s aphasia was interpreted as an impairment of lexical semantics in both language production and comprehension with the grammar module spared. This was seen as evidential basis for the generativist assumption of language modularity.

Such an account was formulated by Grodzinsky and crystallized in the form of the Trace Deletion Hypothesis (first formulated in 1984; and 1986; for later formulations cf. e.g. 1995; 2000; or 2006). The Trace Deletion Hypothesis is a so called representational account of agrammatic (Broca’s) aphasia[8] which means that brain lesions to language areas are taken to partially destroy the representation of linguistic knowledge and language processing in persons with aphasia is thus qualitatively different from neurotypical speakers. This is in contrast to processing accounts which explain aphasic errors as caused by limited processing resources, i.e. the representations as such are intact but the lesions introduce increased noise and stress to the system which decreases the efficiency of the “computational power” allocated for language processing and results in more errors. Grodzin-

8 The discussions that emerged after Caramazza and Zurif’s 1976 influential paper have mostly revolved around agrammatism and agrammatic aphasia which, for the most part, overlaps with Broca’s aphasia.

sky's central claim is that people with agrammatic aphasia lose the ability to represent traces, i.e. phonologically empty elements that are claimed to be part of linguistic structure in generative syntax. For instance, the subject of the sentence *The glass cracked* is underlyingly interpreted as a thematic object of the verb which is moved to subject position during the generation of the surface form and the sentence is in fact taken to have underlyingly the form *The glass*$_i$ *cracked* t_i.[9] This movement leaves behind a trace t_i which is processed as part of the structure and cues the interpretation of the sentence. On Grodzinky's account, traces are deleted and thus not part of the representation of syntactic structures in persons with agrammatic aphasia. For instance, Grodzinsky claims that semantically reversible passive structures, such as *The cat was chased by the girl* are interpreted in agrammatic aphasia as illustrated on the right hand side in Figure 1 as opposed to neurotypical processing with the trace intact shown on the left hand side. Due to missing trace, the subject NP competes for agenthood assignment with the *by* phrase, because both of these syntactic structures are correlated with agenthood in English syntax and both referents are semantically plausible agents.

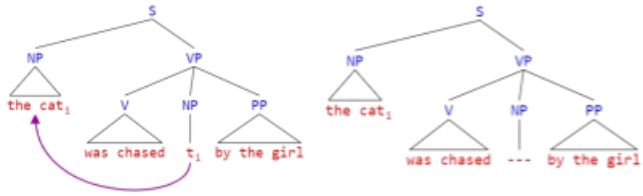

Figure 1: Neurotypical and agrammatic processing of passive sentences under Grodzinsky's model.

The Trace Deletion Hypothesis was later complemented with an account of agrammatic production called the Tree Pruning Hypothesis. Friedmann and Grodzinsky (1997) tested in a case study a speaker with Broca's aphasia on a range of language production tasks and found that the participant made substantially more errors in tense as opposed to agreement in verbal inflection. Similar results were later reported for groups of speakers with agrammatic aphasia by Friedmann (2001; 2006). Friedmann and Grodzinsky interpreted these observations on the basis of the relative position of different inflectional nodes in the syntactic tree such that the agreement node, which is lower in the hierarchical structure according to generative theory, is spared in agrammatic aphasia while the tense node (and all the nodes above) is not available in language production which results in omission and/or substitution errors.

While the Trace Deletion and the Tree Pruning hypothesis claimed to provide evidence for the generative model of language (cf. the title of Friedmann's 2001 paper "Agrammatism and the Psychological Reality of the Syntactic Tree"), this representational account has been controversial since its publication. The approach has been subject to criticism both from structuralist and functionalist perspectives. Berndt and colleagues (1996) performed a meta-analysis of 15 picture-matching studies of comprehension of reversible passive sentences in persons with agrammatic aphasia and found a proportion of the results of these studies to be inconsistent with the predictions of the Trace Deletion Hypothesis and concluded that Grodzinsky's unitary account cannot explain the variation ob-

9 In this notation, *t* represents the trace and the subscript co-indexes the trace with the moved phrase.

served in agrammatic aphasia. A series of studies of online processing using eye-tracking (Dickey, Choy & Thompson 2007; Thompson & Choy 2009; Choy & Thompson 2010) has found that eye movement patterns of people with agrammatic aphasia compare to neurotypical speakers in that there are signs of gap processing in both groups even in structures that are comprehended incorrectly by individuals with aphasia, suggesting that gaps are not deleted.[10] These and other similar findings (see Avrutin 2001; or Bastiaanse & Jonkers 2012 for an overview of linguistic approaches to agrammatic comprehension) has led to the formulation of many alternative accounts that employ the descriptive and conceptual apparatus of the generative theory but argue for a processing account of the deficit. A representative example is the Argument Structure Complexity Hypothesis formulated by Thomspon (2003) who describes that the performance of persons with Broca's aphasia on verb retrieval tasks is correlated with the number of arguments in a verb's argument structure. She also found that unaccusative intransitive sentences are more difficult than unergative intransitive sentences. Thompson argues that the more complex the argument structure of the verb is the more difficult the sentence is to process in agrammatic aphasia due to higher cognitive cost of the processing of such structures.

A very different, but also influential account was proposed by Kolk and collaborators (Kolk & Heeschen 1992; Kolk 1995; Hartsuiker & Kolk 1998; de Roo, Kolk & Hofstede 2003). They proposed that the features characteristic of agrammatic production are similar to elliptical structures of neurotypical speech and that agrammatic speech is not a symptom of aphasia but rather an adaptation strategy that helps speakers deliver the intended message in the context of reduced processing resources. Kolk and Heeschen (1992) compare in this regard speakers with Broca's aphasia to Wernicke's aphasia and argue that the substitutional errors and verbose, but incoherent production characteristic of Wernicke's aphasia is precisely a result of such a computational overload, because speakers with this aphasia type are not capable to adapt their production to the limited resources. Kolk (1995) offers a unifying account of agrammatic production and comprehension via a slowdown of syntactic computation and a rapid decay of results of syntactic processing.

From a more general perspective of theoretical linguistics, the (representational) accounts based on generative syntax have been criticized from cognitive-functionalist positions by Bates and collaborators. This group of studies applied the Competition Model to the study of aphasic language which is, in a way, one of the first examples of a usage-based approach in linguistic aphasiology. The Competition Model (Bates & MacWhinney 1987; Bates & MacWhinney 1989; see also Presson & MacWhinney 2011 for a discussion of language disorders from the perspective of the Competition Model) is a cross-linguistically informed, connectionist model of sentence processing based on the idea that there are different semantic and morphosyntactic cues that guide the interpretation of sentences.[11] These cues have a certain validity, i.e. the informativeness of a given cue. Cue validity is a product of cue availability and reliability, which are roughly defined as the proportion of cases in which a given cue appears in the sentence and the proportion of cases in which the cue leads to the correct interpretation. Both of these measures can be computed from a corpus of sentences in a given language. For instance, preverbal position as a cue for subject- and agenthood is both highly available and reliable in English, whereas in Czech this cue has a lower validity and case and agree-

10 Gap is, in simplified terms, a less theory-dependent equivalent of trace.

11 For a brief characteristic of connectionist models of language representation and processing, see, e.g. Joanisse and McClelland (2015).

ment have more value. Cues can compete or conspire with one another, e.g. both animacy and preverbal position are strong cues for agenthood in English, but may come to conflict in sentences such as *The ball hit the man*. It is apparent that the Competition Model is conceptually close to usage-based models in that it operates with probabilistically defined cues rather than rules and that it is grounded in performance and also emergentist in that the cues and their validity are acquired from linguistic experience rather than innate (e.g. MacWhinney 2015). Bates and collaborators explain agrammatic error patterns in terms of limited processing resources. They claim that the processing of morphology, rather than syntactic structure, is selectively vulnerable in such conditions (see also Boye & Bastiaanse 2018, discussed below) and that cues with higher validity values are more resistant.

Wulfeck et al. (1991) demonstrated such a pattern in a comparison of English and Italian speakers with Broca's aphasia in a series of grammaticality judgment tasks. Speakers of both languages were on the whole more successful in identifying word order violations than agreement errors. However, the English speakers performed better on word order, while the Italian speakers were comparatively more sensitive to agreement errors. This pattern is predicted by the fact that word order has higher cue validity than agreement in English and the opposite is true for Italian. Furthermore, Bates et al. (1987) reported error patterns consistent with the assumptions of the Competition Model that were similar to Broca's aphasia in speakers with other aphasia types and even in non-neurological patients and, moreover, Blackwell and Bates (1995) show that such profiles even occur in neurotypical speakers under cognitive strain, suggesting that, rather than a specific syndrome connected to defined lesion sites, agrammatic errors arise in various populations as a function of diminished cognitive resources and the typological profile of a given language. In a paper with a broader scope, Bates and Goodman (1997) present arguments against a sharp distinction between grammar and the lexicon using data from child language, neurotypical online processing measures, and aphasia. Building on previous research, including the studies mentioned here, they argue against the double dissociation of grammar and semantics in Broca's and Wernicke's aphasia based on the observation that both agrammatic and anomic errors occur across aphasia types. Furthermore, not only agrammatic but also paragrammatic errors occur in Broca's aphasia and the proportion of these two error types is closely correlated to the typological profile of particular languages.

In a more recent example, aphasic data was used to support Boye and Harder (2012)'s *ProGram* theory of the division between grammatical and lexical items in language. This usage-based account adopts the general framework of cognitive functional linguistics but rejects the widely-held idea that there is not a substantial difference between lexical and functional items. Instead, Boye and Harder argue that there is a distinction which reflects the differences in discourse status of these two classes of items. While lexical items are discoursively foregrounded because they carry the core of a message, grammatical items are backgrounded. Boye and Harder derive implications for aphasia research. from this distinction Under their model, grammatical items should be more susceptible to impairments of the capacity to combine pieces of information into more complex units as they must necessarily combine with lexical items. Furthermore, under the assumption of reduced resources, the *ProGram* theory predicts that available processing resources will be allocated primarily to lexical items as discursively more important parts of the message, which increases the probability of omission or substitution errors for grammatical items. These prediction have been tested on lexical and grammatical prepositions (Messerschmidt et al. 2018; Martínez-Ferreiro et al. 2018), pronouns

(Ishkhanyan et al. 2017; Martínez-Ferreiro et al. 2018), and verbs (Boye & Bastiaanse 2018). For instance, Boye and Bastiaanse (2018) investigated the distinction between grammatical and lexical uses of Dutch modal verbs and the verb *hebben* 'have'.[12] They tested a group of 18 speakers with agrammatic Broca's aphasia and ten speakers with fluent aphasia (Wernicke's or anomic) with the prediction that grammatical items are more affected in agrammatic aphasia, whereas lexical items are more affected in the fluent aphasias. They elicited semi-spontaneous speech and coded the verbs in question as used either grammatically, or lexically. The results were consistent with the predictions, with more lexical uses in the agrammatic group and more grammatical uses in the fluent group.

In conclusion, I have argued that linguistics and aphasiology should cooperate more closely for mutual benefit and I have shown this in three examples. First, the Trace Deletion Hypothesis has employed generative theory to explain the symptoms of agrammatism and used this data to support the generative model of language representation and processing. Secondly, the account based on the Competition Model shows how a linguistic theory that integrates language with domain general processes approaches the explanation of agrammatism and how the predictions generated by this model for language processing in general can be tested on aphasic data. Thirdly, the studies arguing for Boye and Harder's model show how aphasic data can serve as an evidential basis for a usage-based model of language. The usage-based studies also show that the narrowly syntactic account offered by researchers working within the generative paradigm formulates too strong predictions about language in aphasia and that the cognitive-functional framework offers an account with more explanatory power that also generalizes beyond aphasia. Section 2.5 follows up on this with a more detailed discussion of the benefits of usage-based linguistics for aphasia research.

The studies within the Competition Model framework presented in this section also demonstrate the importance of linguistic diversity in aphasia research, which warrants a short comment. A vast majority of studies in linguistic aphasiology has been conducted on English in particular and western European languages in general. Beveridge and Bak (2011) analyzed the languages of aphasia research from the period of 2000 to 2009 and showed that 62 % of published studies were based on English and that Germanic and Romance languages accounted for 89 % of all papers in their sample. There has been some awareness in aphasiology toward the need for diversity in this respect (cf. Menn & Obler 1990's cross-linguistic compendium of agrammatic aphasia; or Paradis 2001 which is a volume that contains the papers published in a special issues of Journal of Neurolinguistics on the manifestation of aphasia symptoms in different languages), however the dominance of English has not only been reduced, Beveridge and Bak conclude that it has been increasing. Similar to general and theoretical linguistics as well as other linguistic disciplines, this is a problem, given the, in some respects, unusual typological profile of English caused by language external historical developments (McWhorter 2002). As will be apparent in the next section in which I briefly discuss existing linguistic accounts of aphasia in Czech, the language fits this pattern perfectly, falling very much in the 'under-researched' category.

12 Two primary diagnostic criteria are used in ProGram to distinguish between lexical and grammatical use: only lexical items can stand alone (as in a single word utterance) and be focalized (Boye & Harder 2012: 13–18).

2.3 Aphasia in Czech (linguistics)

Before presenting existing linguistic research on aphasia, I would like to offer a brief overview of the state of speech language therapy in Czechia. People with aphasia have access to speech language therapists (SLT) both as in- and outpatients, as part of the recovery and rehabilitation scheme (for a brief overview, cf. also Lehečková 2012). Some of the tools used in SLT practices for assessment include the Czech adaptation of the Mississippi Aphasia Screening Test (Košťálová et al. 2008; Košťálová 2012; for information regarding the original test cf. Nakase-Thompson et al. 2005); two original assessment tests developed specifically for Czech are also available. Vyšetření fatických funkcí (Phatic functions assessment, Cséfalvay, Klimešová & Košťálová 2003) tests an array of components of language and provides clinicians with a complex profile of the state of the linguistic capacities and aphasia type, based on the Boston classification system (Goodglass & Kaplan 1972b). Dotazník funkcionální komunikace (Fucntional communication questionnaire, Košťálová et al. 2015) is an assessment of pragmatic competence and functional communication. It tests the ability of the examinee to use language in different situational contexts, e.g. filling in a form or talking to an emergency call operator. As far as therapy tools are concerned, different materials are used. Apart from educational materials for L1 and L2 speakers, there are also some tools designed specifically for aphasia rehabilitation. One of these is AfaSlovník (Macková, Macko & Hanibalová 2020) which is a web-based application focused on lexical retrieval training that has a patient and a therapist module which allows SLTs to monitor the performance of their clients and to personalize the tasks according to clients' needs. Importantly, this application is based on corpus and category norm data (Chromý, Diatka & Džupová 2015). It might be said that even though Czech persons with aphasia do have access to therapy, there are relatively few (adapted or original) standardized assessment tools and a lack of therapy materials developed specifically for the needs of people with aphasia. The Czech situation is furthermore characterized by a lack of cooperation between SLTs and linguists, most SLT only receive a rudimentary linguistic training within the scope of a secondary education grammar book.[13] As will become apparent below, there has similarly been little interest in aphasia in Czech linguistics.

Despite the fact that Czech entered aphasia research very early on in Pick's research who worked with Czech-German bilingual patients (Pick 1913; Eling 1994), Czech is a relatively under-researched language with regards to linguistic research of aphasia. Helena Lehečková played a pioneering role in the introduction of this topic to Czech linguistics. Lehečková has published several papers that address the relationship of aphasiology and linguistics, arguing for a closer collaboration between the disciplines and drawing attention of the Czech linguistics community to the ways in which language in aphasia can be used in model building in theoretical and experimental linguistics (2016; 2009; 1985). Lehečková is also the author of the first linguistic descriptions of the symptoms that occur in Czech speakers with aphasia. She analyzed the recordings of connected speech of 35 persons with aphasia that contain a picture description task, a fairytale retelling task, and talk about the participants' illness. Lehečková has mostly focused on agrammatism in Czech and the ways in which this symptom interacts with the typological profile of a given language as well as some crude psycholinguistic measures. In line with the views of other authors (e.g. Badecker & Caramazza 1985; Heeschen 1985; or Goodglass & Menn 1985), Lehečková argues that there is no clear cut distinction between what has traditionally been defined as agrammatic and paragrammatic, particularly

13 It should be noted that there is no official data to support this claim. It is based on informal interactions with SLTs.

on the grounds of crosslinguistic differences in aphasic symptoms. Such proposals are primarily based on the observation that omissions of grammatical affixes that would result in a nonword lexical stem are unattested in aphasic speech.[14] In her analyses of agrammatic errors in Czech (2009; 2001), she focused on a group of 17 speakers with motor aphasia[15] and described the loci and types of errors produced in connected speech. More specifically, she analyzed verbal and nominal inflectional morphology and compared the rates and directions of paragrammatic substitution errors therein (e.g. singular instead of target plural, masculine instead of target feminine, etc.). She related (informally, without any statistical analysis) the patterns in the data to the order of acquisition and frequency of use based on a frequency dictionary of Czech (based on Těšitelová 1985) and found that the later in linguistic development a category is acquired the higher the error rate is and that substitution errors overwhelmingly follow the frequencies of the categories, e.g. when they made an error, the speakers with motor aphasia in Lehečková's sample by and large produced the nominative forms instead of context appropriate cases. She concluded that Czech speakers tend to "default" to the masculine singular nominative in nominal inflection and to the 3rd person singular present active indicative imperfective forms in verbal inflection in aphasia (1988). Lehečková argued that the results are in line with Jakobson's reversal hypothesis (1980 see also section 2.2 above). It is worth noting that these findings are generally in line with the assumptions of the usage-based approaches. However, as discussed in detail in section 2.4, the effects of frequency and the predictions derived therefrom are much more complex and nuanced.

The other major contribution to the characterization of aphasia in Czech is a series of papers by Flanderková & colleagues who tested several predictions and findings published for other languages. Hudousková et al. (2014) tested the predictions of Grodzinky's Trace Deletion Hypothesis (e.g. 2000; see also 2.2). They tested six participants with agrammatic Broca's aphasia on active and semantically reversible and irreversible passive sentences in a picture selection task. They found that participants with aphasia performed virtually at ceiling level for active sentences and that semantically reversible passive sentences were on the whole more difficult than irreversible sentences. However, the error rates were above chance level, from which they conclude that the Trace Deletion Hypothesis does not hold for Czech. Apart from a low number of participants and experimental stimuli which is quite typical in the field, it is worth noting that the authors did not collect acceptability judgment rates for the experimental stimuli, nor did they retrieve corpus frequencies for active and passive uses of the stimulus verbs, which is a variable that may play a role under the usage-based model.

In the largest experimental study concerned with the performance of persons with aphasia on various linguistic tasks available for Czech, Flanderková (2019) presents the results of four sets of experiments motivated by various findings from the literature. Flanderková worked with a group of 15

14 For instance, the Czech words *žena* 'woman' and *duše* 'soul' are morphologically complex and analyzable into the roots *žen-* and *duš-* and grammatical suffixes *-a* and *-e*, marking singular nominative word forms in feminine words. While the form *žen* is a free morpheme in Czech (it is the form of the plural genitive), *duš-* is a bound morpheme. It has long been established that bare stems that are bound morphemes are never produced by speakers with aphasia. It follows from this that morphologically rich languages with inflectional morphology tend to have a much lower proportion of omitted grammatical morphemes (for one of the first such accounts cf. Slobin 1991 and his analysis of the manifestations of aphasia in Turkish). It should be noted that e.g. Hudousková et al. (2014) claim that the prediction in agrammatic aphasia is for lemmas such as *postel* 'bed' that do not possess a grammatical morpheme in their citation form to "default" to this form which effectively is an omission error.

15 No clinical details are provided in the papers.

Czech speakers with different types of aphasia. One of the results that lend support to the importance of research on diverse languages are Flanderková's experiments focused on the role of argument structure in verb retrieval. She tested the predictions of Thompson's Argument Structure Complexity Hypothesis (2003) using confrontation naming, picture matching, and a task where the participants categorized verbs according to valency. Flanderková failed to find a conclusive effect of valency (cf. the results of the analysis in section 4.3.1). In another "typologically informed" experiment, Flanderková tested the sensitivity of speakers with aphasia to information structure driven word order differences in Czech by having the participants choose a context appropriate continuation of a miniature discourse. She found that both fluent and nonfluent speakers with aphasia performed similarly and above chance level (60 % of correct responses). This is in line with previous research on discourse processing in aphasia that showed that discourse level phenomena tend to be relatively spared in aphasia.[16] It should be noted that both the number of participants and the number of stimuli in Flanderková's experiments were relatively small. The results of the verb retrieval experiments rely on the framework of the Functional Generative Description and the entries provided in a valency dictionary of Czech verbs, *VALLEX* (Lopatková et al. 2020). The entries in VALLEX are based on corpus data and the descriptions are based on the Functional Generative Description framework (e.g. Panevová 1980), which is, in some respects, a formal theory with discrete yes-no categories, which may understandably have also influenced the results.

In conclusion, the language behavior or Czech speakers with aphasia has only attracted a limited interest in the linguistic community. While there are important contributions that provide some basic characteristics of the manifestation of aphasia in Czech based both on observational and experimental data, more research is needed for a deeper understanding of aphasic Czech. One salient gap that the present work is trying to contribute to fill is the lack of discourse level and connected speech oriented studies. Note also that the research presented in this sections has not reflected on recent developments in linguistics. The studies by Flanderková and collaborators were modeled on various findings from the literature that are more or less grounded in the generative theory. These were combined with some of the theoretical constructs of the Prague functional-structuralist linguistic framework which also formed a frame of reference for Lehečková's findings.

As a concluding remark to this section, it should be noted that I have only focused on linguistic descriptions of aphasic data. That is not to say that there is a complete lack of aphasia research in Czechia. Practicing SLTs and graduates of SLT programs indeed publish papers, most typically case studies, that discuss various facets of aphasia, primarily approaches and strategies of rehabilitation. However, there is an almost complete lack of communication and cooperation between SLTs and linguists and this lack in turn, in my opinion, hinders the development of focused, more sensitive, and reliable assessment and therapy tools.

2.4 Usage-based linguistics: a brief introduction

The present work adopts usage-based construction grammar as a model of language representation and processing and uses this conceptual framework to describe some aspects of connected speech

16 Typical problems in discourse production and comprehension involve referential tracking problems (e.g. introduction of pronominal referents with no clear preceding antecedents) and the processing of implicit information (e.g. Ulatowska, Allard & Chapman 1990).

production of Czech speakers with aphasia. In this section, I describe the foundational assumptions and concepts of usage-based construction grammar.

The term *usage-based model* was first used by Langacker (1988) and refers to the general idea that usage matters in language. In more specific terms, usage-based construction grammar understands language knowledge as a network of constructions that is shaped by communicative pressures and cognitive capacities via usage. It is part of a family of functionalist and cognitive linguistic approaches to language and shares with them, among many other assumptions and concepts, the cognitive commitment, i.e. the idea that generalizations and explanations in linguistics should not be cut off from other disciplines, but rather integrate the findings from other fields of cognitive science (Lakoff 1990; Divjak, Levshina & Klavan 2016; for recent representative studies see Schmid 2016; Divjak 2019; Goldberg 2019; or Diessel 2020).

The term usage-based refers to the idea that language structure is directly affected by usage and that structure and systematicity in fact emerge from usage.[17,18] Individual language users possess a representation of language and grammar which is directly based on their linguistic experience which consequently modulates mental grammars throughout lifespan. Several important points follow from this general assumption. Firstly, performance factors that have been viewed as secondary and unimportant in the study of language competence in generative grammar are an integral part of the usage-based model of language because these forces directly shape language structure. To provide a very simple example, take the irregular English plural *mice* which has not been regularized due to the conserving effect of high frequency, which ultimately reflects the facts of the world people talk about more (cf. e.g. Bybee 2006; or 2010).

Secondly, the use of large amounts of authentic language data and robust empirical methods is an integral part of the approach. In other words, usage-based analyses make use of large corpora of written, spoken, and signed language but also experimental methods and, increasingly, the combination thereof (cf. a representative example of studies reprinted in Janda 2013; and Divjak, Dąbrowska & Arppe 2016 for the discussion of corpus and experimental data). For instance, Klavan (2012) analyzed the factors favoring the selection of an adpositional phrase or the locative case in Estonian and used a rating task that was designed based on a previous corpus analysis to validate the corpus data and compare it with behavioral data (see also Klavan, Pilvik & Uiboaed 2015; Klavan & Divjak 2016; Klavan 2017; or Klavan & Veismann 2017). This requirement appeared originally as a reaction to the use of introspection and informal acceptability judgments of made up examples typical for (early) generative analyses (cf. Dąbrowska 2010 for arguments against introspection; or Gibbs 2006; and Willems 2012 for a discussion of the merits and disadvantages of introspection within the context of Cognitive linguistics).

Thirdly, the idea that mental grammars remain dynamic and may be "updated" during each and every usage event implies that the language knowledge of individual members of a given language community need not and, in fact, should not be identical. For example, Street and Dąbrowska

17 Language is in this respect viewed as a complex adaptive system (Beckner et al. 2009), i.e. a system that emerges from localized interactions of agents who employ different strategies to achieve their (communicative) goals and adapt these strategies for future interactions based on the outcomes of past experience.

18 The following is based on the claims presented e.g. in Tummers and colleagues (2005), Bybee (2010), Taylor (2012), Diessel (2015), and Goldberg (2019).

(2010) compared a group of postgraduate students and participants with non-graduate education and in a picture selection task testing the comprehension of the passive and the universal quantifier construction in English. While the postgraduate participants performed at ceiling levels on all sentence types, the lower education level group performed at ceiling only in the control condition (simple transitive sentences). This difference can be explained by different language experience in both groups (see also Dąbrowska 2012 for a review of studies demonstrating differences in individual mental grammars; or Verhagen 2020; and Barking, Backus & Mos 2022 for recent examples of studies with similar conclusions). Related to this, the usage-based model takes the position that the major driving force in language change are adult users.

Lastly, the usage-based approach refuses claims of (strong) innateness. Rather than assuming that certain linguistic abilities are innate as claimed in generative grammar, usage-based linguistics argues that there are certain aspects of neurocognition that are fine tuned to language (e.g. in auditory processing) but that linguistic categories or other aspects of language structure are not innate but rather emerge during acquisition thanks to domain general processes, such as statistical learning, chunking, categorization, or analogy (Thompson & Newport 2007; Christiansen & Chater 2015). In a representative study, Dąbrowska and Lieven (2005) analyzed the development of syntax in two English children based on transcripts of spontaneous speech. They argued convincingly that the data was better characterized not in terms of innate abstract categories but rather in the usage-based perspective, with children arriving at syntactically complex structures and abstract categories by employing two simple operations (juxtaposition and superimposition) on an inventory of lexically specific constructions.

The basic unit of language representation in usage-based construction grammar is a construction. Constructions in the technical sense are defined as direct pairings of form and function (meaning) with function potentially integrating semantic, contextual, and pragmatic meanings as well as world knowledge (Goldberg 1995; Croft 2001; Fried & Östman 2004). Constructions have different levels of inner complexity and schematicity. A construction may comprise a single word or several words with varying levels of syntactic complexity, while schematicity ranges from lexically specific constructions with limited compositionality to fully abstract schemas. Thus, *pes* ‘dog’ is a single word lexical construction, *pod psa* ‘under the weather’ is a fully fixed lexically filled non-compositional construction which, however, is also an instantiation of a more general pattern [pod N-ACC] which in turn is an instance of a more abstract construction [[SPATIAL PREPOSITION] [N-ACC]$_{\text{GOAL}}$]. *Starého psa novým kouskům nenaučíš* ‘you can’t teach an old dog new tricks’ is an instance of an idiomatic construction which allows for some lexical and/or morphosyntactic variation,[19] and, finally, *Jiří vyvenčil psa* ‘Jiří walked the dog’ is a largely compositional instantiation of the transitive construction [N-NOM]$_{\text{AGENT}}$ [V] [N-ACC]$_{\text{PATIENT}}$. Word internal structure is also conceived of in constructional terms (e.g. Booij 2010a; 2010b; or 2018). Thus, the Czech place/container nouns of the type *botník* ‘shoe cabinet (lit. shoe place)’, *kurník* ‘hen house (lit. hen place)’, or *rybník* ‘pond (lit. fish place)’ can be rendered as instantiations of an abstract pattern [[X]$_{\text{N}}$ník]$_{\text{N}}$ with the meaning component

19 Some of such variations attested in syn v9 include: *Starého komunistického psa holt novým kouskům nenaučíš.* ‘You just can’t teach an old commie dog new tricks’, *Starý pes umí nové kousky* ‘An old dog has learned new tricks’, *naši lídři připomínají staré psy, kteří nemají snahu učit se novým kouskům.* ‘our leaders resemble old dogs that show no effort to learn new tricks’, or *Staří psi předvádějí staré kousky* ‘Old dogs perform old tricks’. These variations show that the use of the “skeletal” elements *starý* ‘old’, *pes* ‘dog’, *nové* ‘new’, *kousky* ‘tricks’, and *naučit* ‘teach’, or paradigmatically related words suffices to evoke this idiom (cf. Fried 2013 for a similar analysis).

[[CONTENT]ník]$_{CONTAINER}$. Even abstract grammatical constructions such as various argument structure constructions retain some semantic value which constributes to the meaning of the construction. This is why a speaker of Czech would probably not have problems understanding the sentence *Kouzelník zmizel králíka* 'The magician disappeared the rabbit' even though it would be considered ungrammatical or borderline acceptable, because the verb *zmizet* 'disappear' in normally used only intransitively in Czech. Similarly, a speaker's experience with the Czech transitive construction, together with world knowledge, contributes to the fact that a sentence such as *Páže zraní laň* will be without additional context arguably interpreted as 'The squire will hurt the doe' with *páže* 'squire' being the subject and agent of the sentence despite the fact that the opposite is also plausible due to non-rigid word order and case syncretism of the nouns. This assumption is closely related to the fact that grammar and the lexicon are viewed not as two distinct categories but rather as a cline with prototypical grammatical morphemes on the one end and prototypical lexical morphemes on the other (but cf. Boye & Harder 2012 who argue against this view).

Constructions are represented in language knowledge in a network (of networks) in a multidimensional conceptual space. The networks are organized based on overlaps of function and/or form (Fillmore 1988; Goldberg 2019; Diessel 2020). Language processing operates on this network of constructions. During every usage event, language users compare the incoming linguistic data and context (linguistic and situational) against their previous language experience (Christiansen & Chater 2015; Christiansen & Chater 2016; Chater & Christiansen 2016). Based on this experience, they select the most plausible interpretation of the linguistic signal. Similarly, during language production, such signal is selected based on previous usage events that best fits the context and the goals that the speaker wishes to achieve in interaction. This accumulating language experience, combined with memory processes and the ability to recognize similarities and apply analogy lead to the clustering of constructions with function and/or form overlap and gives rise to generalizations (Perek & Goldberg 2017; Goldberg 2019). Ultimately, then, it might be argued that categorization is at the very heart of language processing under the usage-based model. It follows from these assumptions that categories in usage-based construction grammar are prototypical in nature (in the sense of Rosch 1973), with more central and peripheral members (Taylor 1995; Divjak & Arppe 2013; Ramscar & Port 2015). These clusters and generalizations are also an alternative for grammatical rules that are postulated in structuralist approaches (Dąbrowska 2004; Bybee & McClelland 2005; Abbot-Smith & Tomasello 2006; Kapatsinski 2012). Taylor (2012) provides two metaphors to compare the families of rule-based and usage-based approaches. Language in the structuralist and generativist approaches is compared to a combination of a dictionary and a grammar book: when a speaker wants to produce a form they look up the lemma in the dictionary and apply an appropriate rule based on the grammar book. On the other hand, language in the usage-based perspective is more like a mental corpus with rich representations, contextual information, and a great deal of redundancy. To extend the metaphor a little further, when a speaker plans to produce an utterance we can imagine them to query the corpus to arrive at the most appropriate framing of the utterance based on the query and the distributional characteristics of possible solutions given the context.

Frequency of use plays a crucial role in these processes and has become one of the central and most studied notions in usage-based linguistics (Bybee 2007; Divjak 2019). High frequency and, consequently, repeated processing leads to stronger memory traces and routinization. Highly frequent items are faster and easier to access and have increased probability of phonetic reductions due to

greater overlaps of articulatory gestures. Importantly, different kinds of frequency influence language representation and processing in slightly different ways. Token frequency is responsible for ease of retrieval. A host of studies has demonstrated that highly frequent words are responded to faster in tasks such as word identification or lexical decision (among the first e.g. Howes & Solomon 1951). Type frequency has been demonstrated to play a role in productivity and in language change (regularization, analogical leveling) (e.g. Baayen 2009). Finally, relative frequency of constructional variant or a specific paradigm cell has been demonstrated to play an important role via probabilistic processing (Bod, Hay & Jannedy 2003; Bod 2006; Levshina 2018). This, in simplified terms, means that speakers have access to collocational preferences of words or structures. For instance, Garnsey and colleagues (1997) used English verbs that are biased either toward taking a direct object, or a complement clause, based on the relative frequencies of the complements (e.g. *The neighbor wrote the letter.* vs *The neighbor assumed (that) the letter was fake.*). They investigated the effect of these biases on the resolution of temporarily ambiguous sentences of the type *The neighbor wrote the letter was fake.* and found a processing advantage for the complement clause biased verbs. Furthermore, when two or more words (or morphemes) occur frequently together they may be represented and retrieved as a whole rather than being assembled anew during each usage event (Bybee 2010; Arnon & Cohen Priva 2013; 2015; Verhagen et al. 2018; Onnis & Huettig 2021). All of the assumptions mentioned in this section show that accounts of structural and processing phenomena are primarily made on the basis of domain general cognitive processes that interact with linguistic representations (which are coded in a specific format (Hagoort 2013) in usage-based linguistics rather than language specific mechanisms as argued in generative theory. As will be discussed in the following section, there are good reasons to believe that the mechanisms outlined here can provide more accurate descriptions and analyses of aphasic data.

2.5 Usage-based approaches to aphasia

Apart from the studies grounded in the Competition Model mentioned in section 2.2, linguistic research of aphasia has been until relatively recently dominated by rule-and-lexicon structuralist approaches that have stressed the role of structural complexity in aphasic processing. However, in recent years, analyses of language in aphasia started to emerge that adopt the usage-based approach. These studies have focused on the role of performance factors, mainly frequency and increasingly also the probability of occurrence of inflected word forms and/or cooccurrence of words in constructions,[20] but also age of acquisition effects (e.g. Brysbaert & Ellis 2016), or processing strategies grounded in world knowledge.[21] These studies suggest that the application of the usage-based perspective has the potential to open new avenues of aphasia research and provide accounts of language in aphasia with more explanatory power. This follows from the fact that the usage-based model is performance oriented and draws attention to such phenomena as frequency of use, similar-

20 While the frequency effects of single words have been generally accepted and investigated in aphasia research (cf. Howes 1964; Kittredge et al. 2008; Nozari et al. 2010; or Bastiaanse, Wieling & Wolthuis 2016), there has been a lack of interest in frequency-driven probabilistic processing of multiword units despite the fact that their impact on language representation and processing in neurotypical speakers has been demonstrated repeatedly (Gahl & Menn 2016).

21 This line of research adopts the noisy channel/rational inference approach which stresses the importance of plausibility of interpretation derived from meaning and world knowledge in agrammatic comprehension (cf. Gibson et al. 2016; or Warren, Dickey & Liburd 2017). This approach is outside the scope of the present work and will not be addressed in detail.

ity, world knowledge, textual and situational context, or recency and gives space to inter- as well as intra-categorial differences.

The research of Susanne Gahl was one of the first examples of applying the usage-based framework to aphasic data. In a series of papers, Gahl (2000; 2002; Gahl et al. 2003) investigated the role of verb frame frequency in sentence processing by persons with aphasia. Specifically, Gahl (2000; 2002) focused on the frequency of transitive and intransitive uses and its role in sentence comprehension. Based on corpus frequencies, she identified ten verbs with a transitive bias (e.g. *boil*) and 10 verbs with an intransitive bias (e.g. *melt*) and used these verbs in transitive active, intransitive active, and passive sentences with plausible or implausible semantics (e.g. *The cook melted the butter* vs *The butter melted the cook*). These stimuli were used in a plausibility judgment task with a group of 18 persons with aphasia (six participants with Broca's, four with Wernicke's, two with conduction, and six with anomic aphasia) and five neurotypical speakers. The results confirmed the predictions of Gahl's Lexical Bias Hypothesis that, just like in neurotypical speakers, the performance of persons with aphasia is modulated by relative frequency. Participants made more errors in the sentences that did not match the verb bias. This finding was confirmed in a follow up study with eight persons with aphasia (Gahl et al. 2003) This study also confirmed that the effect of verb bias was present both in participants with fluent and non-fluent aphasia. A similar effect concordant with the Lexical Bias Hypothesis was also discovered in reading, using the self-paced reading method (DeDe 2013a). Furthermore, DeDe (2012; 2013b) found that persons with aphasia make use of these verb biases in the resolution of temporarily ambiguous sentences (e.g. transitive-biased *While the parents watched, the baby sang a song* vs. intransitive-biased *While the parents danced, the baby sand a song*). DiLallo and colleagues (2017) also found this effect in language production. They analyzed data from AphasiaBank (MacWhinney et al. 2011) and found that speakers with aphasia produced more intransitively biased verbs and that agrammatic production was more frequent in contexts of mismatch between sentence structure and verb bias. Anderson (2017) investigated the interplay between verb lemma frequency and the frequency with which a given verb appears in a (argument structure) construction. Anderson used a verbal fluency task that elicited verbs using simple clausal frames composed entirely of function words (e.g. *She [V] at me.* to elicit verbs such as *smile*, *point*, etc.) with a group of four speakers with aphasia and a grammaticality judgment and sentence completion task with a group of 14 persons with aphasia. She found that verb-less, semantically "hollow" argument structure constructions did indeed evoke corresponding verbs. Furthermore, an interaction of lemma frequency and construction frequency was a predictor of the performance of participants with aphasia in both the grammaticality judgment and the sentence completion task.

Hatchard and Lieven (2019) used relative frequency of word forms to describe errors in a corpus of Cinderella retellings by 12 people with aphasia. In contrast to the studies mentioned above, they focused on the production of singular and plural noun forms and conducted a qualitative analysis of agrammatic errors in the data by using the concept of dominance, i.e. the proportion of singular and plural uses of a lemma in a corpus, to see if word form frequency could influence the nature of errors. They found an interesting pattern where several of the participants with aphasia produced the pluralized forms *shoes* or *slippers* in a context where the singular was appropriate. Importantly, this corresponds to corpus data that shows that both of these lemmas are plural-dominant. This study shows how observational data can be approached within the usage-based framework and how

(semi)spontaneous connected speech in combination with the descriptive and analytical apparatus of usage-based linguistics can complement experimental research and generate hypotheses. As will become apparent, the present work adopts a similar approach.

In a related study, Hatchard (2021) analyzed a corpus of Cinderella stories produced by six speakers with aphasia with regards to the production of verbs and verb containing multiword units. In her analysis, Hatchard applied the usage-based construction grammar perspective to derive predictions with regards to connected speech production in aphasia. She analyzed the number and diversity of verbs in her corpus and focused on the length, complexity (number of verbs and clauses), and well-formedness of "strings" (roughly comparable to c-units but designed with the characteristics of aphasic speech in mind). Based on the ressults, Hatchard argued that particular aphasia types or language profiles are better conceived of as points on a continuum rather than discrete categories. She also found that more severe cases of aphasia are characterized by a lack of constructional schematicity and a reliance on lexically specific constructions, possibly represented holistically, with relatively higher frequency (measured as corpus attestation).

A similar finding was also reported by Zimmerer et al. (2018). Zimmerer and colleagues programmed a tool for automated analysis of formulaic language, the Frequency in Language Analysis Tool. This tool extracts the frequency of uni-, bi-, and trigrams from the spoken part of the British National Corpus and computes several indices of formulaicity. In their proof-of-concept comparison of groups (ten participants each) of neurotypical speakers, persons with right hemisphere damage, and persons with fluent and nonfluent aphasia, they show that the production of people with aphasia is characterized by the use of high frequency content words, and by reliance on frequent and strongly collocated bigrams and trigrams. Folowing up on these results, Bruns and collaborators (2021) developed an intervention program that trained participants on 12 high frequency constructions with various interactional functions. The program used structural priming in a word monitoring game that used self-voiced auditory stimuli. The paradigm encouraged participants to apply superimposition by combining trained constructions (such as *I like CDs*) with the generalized schema *[person] like [thing/person/place/activity]* and semantically appropriate complements to create new instances such as *They like the house*. While the results of the case series with five individuals with aphasia were mixed, Bruns and collaborators conclude that some of the participants benefited from this program.

To conclude this section and relate it to the discussion of lingustic theory in the context of aphasia from section 2.2, I would like to take a closer look at a paper by Gahl and Menn (2016) (henceforth GM), which appeared in a special issue of the journal *Aphasiology* focused on the role of frequency and other "usage" variables in aphasia research. GM provide a review of studies of aphasia that employ usage-based approaches and argue for the use of frequency-based probabilistic data in aphasia research for the reasons discussed in this and the previous section.

To show that the usage-based approach has more explanatory power, GM go on to support their argument by reanalyzing the data presented in Bastiaanse et al. (2009) (henceforth, B et al.). B et al.'s paper itself draws on a series of papers that present arguments for the Derived Order Problem Hypothesis. The hypothesis claims that persons with agrammatic aphasia have difficulties with the production of structures that involve movement of either the verb itself or its argument. For example,

based on a sentence elicitation task, Bastiaanse and van Zonneveld (2005) studied differences in the production of sentences involving Dutch verbs that can be used either as transitive or unaccusative, similar to the Czech example 1. They found that the unaccusative frames were much more prone to error in a group of eight speakers with agrammatic Broca's aphasia. Crucially, the generative analysis of such structures assumes that the subjects of unaccusative verbs are structurally objects (and themes) that are moved into an empty subject position (1.b).[22]

1. Transitive and unaccusative use of a verb under generative analysis:

   a. *Kočka rozbila hrnek.* 'The cat broke the mug.'

   b. *Hrnek$_i$ se rozbil t_i.* 'The mug$_i$ broke t_i.'

The 2009 paper amends the analyses by obtaining frequency data for the structures that were investigated in the author's previous research. They conclude that the difficulty of structures involving movement cannot be explained by frequency effects. For instance, their estimates of the frequencies of transitive and unaccusative uses of the verbs used in their experiment are on the whole comparable. They also include the frequencies of transitive and unaccusative uses of the individual verbs and enter this factor in a regression equation, failing to find a significant effect of frequency. They only find an interaction effect of condition (unaccusative vs transitive) and group (agrammatic vs fluent). The agrammatic group was significantly more successful in the transitive condition, while the fluent group was slightly more successful in the unaccusative condition.

GM in their response to B et al.'s results point out that the assumptions behind the model used in B et al.'s analysis are not in line with the predictions made by probabilistic approaches to syntax. The expectation in under this model would be that the probability of success or error would be modulated by the *relative* frequency of the respective constructions. This means that verbs that are used more often in the unaccusative frame would have a higher probability of correct response in the unaccusative condition and vice versa. GM correctly note that this should be expressed in the statistical model as an interaction of condition and frequency. Alternatively, one might also conceptualize the model as follows, *response ~ group + condition + proportion of transitive uses in corpus + cumulative lemma frequency*.[23] GM go on to derive a different variable which is based on group performances across verbs and conditions and demonstrate that there is indeed a noticeable trend in the data that conforms to the predictions of usage-based linguistics.

This approach is illustrated by Figure 2. I used the data presented in Tables C1 and C2 in Bastiaanse et al. (2009). The ratio of correct unaccusative uses to correct transitive uses for individual verbs by the agrammatic group is plotted against the relative frequency of unaccusative uses in B et al.'s corpus. While the reanalysis presented in this section should be taken with a grain of salt because only aggregated data is available, it serves as a good example of the potential that the usage-based perspective brings to aphasia research. The various frequency effects and other similar considerations have the potential to provide better control over experiments, additional explanatory power, and, ultimately, more informed and efficient assessment and therapy tools.

22 It should be noted that the constructivist approaches do not posit movement as a syntactic (transformation) operation but use different analyses.

23 This would make sense as there are marked differences in the frequencies of the lemmas used by B et al.

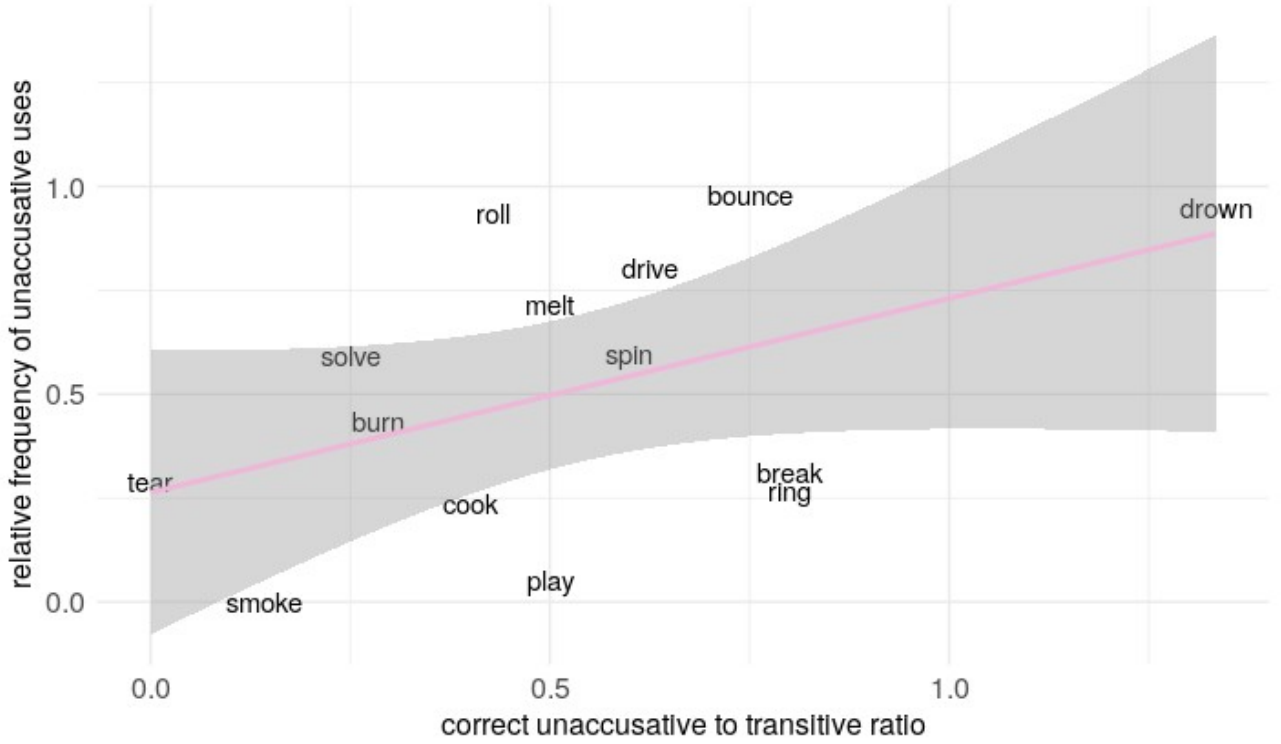

Figure 2: A graph representing a reanalysis of the B et al. study with individual verbs and a trend line. The data is based on aggregated results of speakers with agrammatic aphasia.

2.6 Rationale for a usage-based analysis of aphasic language: A practical introduction

To conclude this section, I provide two more examples from the data that this thesis is concerned with to show how a usage-based analysis proceeds when analyzing language data. Both of these examples come from a single recording of a male speaker (69 years old at time of recording) with chronic conduction aphasia (cf. section 3.3 for further details). The examples in 2 come from a single task and complement nicely the argument presented by Gahl and Menn (2016) to provide a rationale for the analysis of aphasic data from a usage-based perspective.

2. An excerpt from transcript aa3:

1. EXP: *říká(m) H 0.2 jednoduše jenom 0.2 v(o) co tam šlo 0.5 dyby ste mi řekl* 'I'm saying H 0.2 simply just 0.2 what happened there 0.5 if you could tell me'
2. AA3: *H 0.5 vidite 1 todle to to je 1 n- ne 0.5 pes ale 0.2 ř- 0.5* 'H 0.5 you see 1 this it it is 1 n- not 0.5 dog but 0.2 r- 0.5'
3. EXP: *k- 1* 'c- 1'
4. AA3: *a H přitom to normalně používam 0.2 H nebo nepoužívám ale jako vim to 0.2 ryb- ne 0.5 ryby 0.5 byly tady* 'and H just I normally use this 0.2 H or I don't use it but I know it 0.2 fi- no 0.5 fish-PL 0.5 were there'
5. EXP: *tady a tohle je 0.5 pes a 1* 'there and this is 0.5 a dog and 1'
6. AA3: *a* 'and'
7. AA3: *to neni (ryba) 0.5 H 0.5* 'that's not (fish-SG) 0.5 H 0.5'

8. EXP: *ko- 0.5* ‘ca- 0.5’
9. AA3: *k- 0.2 koř- 1 (to) vůbec 1 to si vemu počitač H teda telefon* ‘c- 0.2 car- 1 (that) I don’t 1 I would use a computer H or a phone’
10. EXP: *aha 0.2* ‘aha 0.2’
11. AA3: *a tam si to 0.2 najdu a* ‘and there I 0.2 look it up and’
12. EXP: *najdete 0.2 aha* ‘look it up 0.2 I see’
13. AA3: *tak jako 0.2 a tim H sem 1 dyž se to stane a(le) normálně já nevim (jak) (se) ry- 0.5 H (to) ne tam je (ř) 0.5* ‘well just 0.2 and that way H I’m 1 when it happens b(ut) normally I dunno (how) (it’s) fi- 0.5 H (that) no it has (an r) 0.5’
14. EXP: *tak 0.2 ryba tu je a* ‘so 0.2 a fish is there and’
15. AA3: *no to ryby a 0.2 H todle je 1 r- 1* ‘well it fish-pl and 0.2 H this is 1 f- 1’
16. EXP: *koč- 0.5* ‘ca- 0.5’
17. AA3: *kočky* ‘cats’
18. EXP: *kočka* ‘cat’
19. AA3: *kočky* ‘cats’
20. EXP: *tak* ‘yes’
21. AA3. kočky 0.2 H 0.2 kočka viděl toho (no) (to) 0.2 a sež- sežra- 0.2 vytáh 0.2 toho ryby 0.2 toho ry- rybu 0.5 a sežral jí ‘cats 0.2 H 0.2 cat saw-M the-M.ACC (well) (that) 0.2 and dev- devou- 0.2 pulled-M out 0.2 the-M.ACC fish-PL 0.2 the-M.ACC fi- fish-SG 0.5 and devoured-M it-F.ACC

The two phenomena I would like to focus on here are both examples of paragrammatic errors. The first one appears in turns 4, 5, and 21 where the participant erroneously uses a plural form of the lemma *ryba* ‘fish’ (*ryby)*. The second error, or, rather, a cluster of errors appears again in turn 21 where the participant produces a string of agreement errors. The past tense verb forms *viděl* ‘saw’, *vytáh* ‘pulled out’, and *sežral* ‘devoured’ as well as the demonstrative pronoun *toho* the-M.ACC’ all have masculine agreement morphology, whereas the controllers of these forms are in all cases feminine (*kočka* ‘cat’ and *ryba* ‘fish’ respectively), such that the target grammatical form of the utterance in turn 21 would be *kočka viděla a vytáhla tu rybu a sežrala jí* ‘the cat saw-F and pulled-F the fish-SG out and devoured-F it’.

While the gender mismatch would probably be explained as a reversal to a category that is unmarked, such an explanation would not be possible in the case of the number mismatch, as plurals would be by default considered more marked than singulars under a structuralist analysis. Apart from the fact that markedness is a concept with a host of theory dependent definitions (cf. Haspelmath 2006), it would fail to explain the number related paragrammatism. On the other hand, usage-based linguistics provides a framework that readily offers an explanation of both of the paragrammatisms. As shown below, the explanation is essentially grounded in the assumption that language users track the probabilities of different linguistic phenomena across different contexts.

Let us start with the number mismatch. There is a number of studies in the psycholinguistic literature that suggest an effect of number dominance, i.e. a high relative frequency of either the singular or plural word forms, in word processing (cf. e.g. Baayen et al. 2003; or Beyersmann et al. 2015; for the effects of number dominance in aphasia cf. Biedermann et al. 2012; or Biedermann et al. 2018). Furthermore, under the usage-based model, highly frequent word forms would be expected to have a relatively strong memory trace and thus be retrieved as wholes rather than assembled *de novo*, as would be expected by some models of inflectional morphology (e.g. Lukatela et al. 1980).

In a crude and simplified model of representation of inflectional morphology which draws on Booij's Construction Morphology (2010a), the word *ryba* is an instantiation of an abstract pattern [[R]-a] which is associated with the meaning of singular nominative of (mostly) feminine nouns. This pattern is connected with the plural scheme [[R]-y]. A plural form, such as *ryby* might be directly represented as a fully specific construct if its frequency is considerably high, or it may be produced by taking the form *ryba* and combining it with the plural pattern, employing the partially filled word-level construction and analogy with nouns with similar forms and/or meanings.[24] The usage-based approach would predict that singular, plural, or both forms may be directly represented in one's linguistic knowledge, depending on linguistic experience with the respective word forms. In the context of language impairments, it would then be predicted that the word forms that will be easier to retrieve and more resistant to error will be the ones that can be assumed to be stored directly and not constructed by means of an abstract schema. In lemmas that are more frequent in plural than in singular the probability that the speaker "misfires" and retrieves such a word form even in inappropriate contexts rises.

Table 4 shows the distribution of individual paradigmantic cells of the lemma *ryba* 'fish' in a corpus of spoken Czech (2017) (henceforth ORAL) with relative frequencies of the individual combinations. As can be seen, the plural accusative word form *ryby*, i.e. the one produced in turns 4, 5, and 21 in 2 above, is the most frequent, constituting more than a quarter of the total occurrences. When combined with the syncretic plural nominative form, it constitutes almost 50 % of all occurrences. The lemma is plural dominant, with 61.69 % of plural uses. As can be further seen in Figure 3, just three out of the 10 theoretically possible word forms (regardless of number-case values) make up over 80 % of the lemma's occurrences. It should be noted that these include both the target word forms from example 2 *ryba* and *rybu*. The produced paragrammatic word form *ryby* is more than twice as frequent as the respective targets.

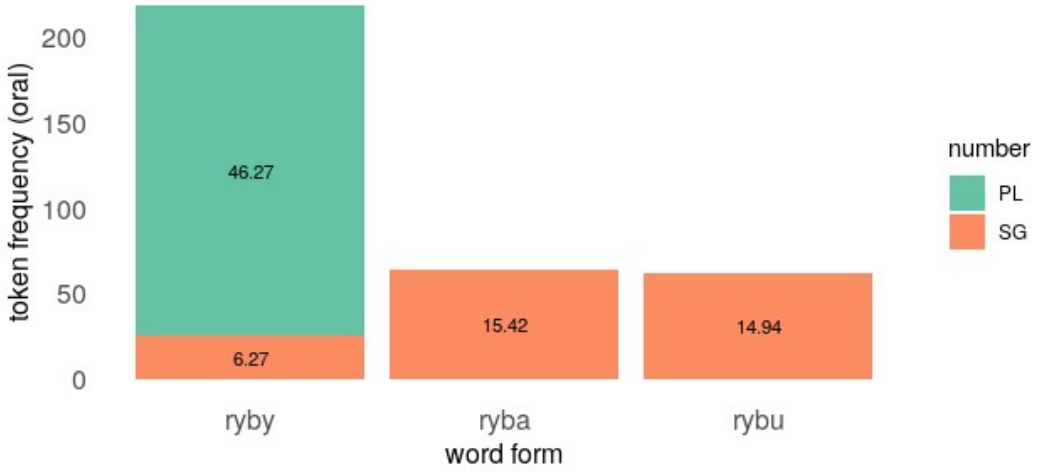

Figure 3: Token frequencies of selected word form of the lemma ryba. Numbers on the columns indicate percentages.

24 This would, for instance, rule out the hypothetical form *kolegy* as plural of *kolega* 'colleague', because there is a shared form with the pattern in question, but the meaning differs in that it refers to a male person whereas the lemmas that do use the plural pattern [[R]-y] and refer to human(oid) beings always denote female persons.

word form	number	case	frequency (pmw)	relative frequency
ryby	PL	ACC	21.98	0.2843
ryby	PL	NOM	13.78	0.1783
ryba	SG	NOM	11.92	0.1542
rybu	SG	ACC	11.55	0.1494
ryb	PL	GEN	6.71	0.0867
ryby	SG	GEN	4.84	0.0627
rybách	PL	LOC	2.42	0.0313
rybama	PL	INS	1.68	0.0217
rybám	PL	DAT	1.12	0.0145
rybě	SG	LOC	0.75	0.0096
rybou	SG	INS	0.56	0.0072

Table 4: Frequency distribution of paradigmatic cells of the lemma ryba 'fish'; frequency is from the corpus oral.

If we assume that individual forms that have significant overlaps compete in language processing (e.g. Goldberg 2019), it follows quite naturally that the number mismatched, yet highly frequent word form might be retrieved under the assumption of limited processing resources. In this respect, the pluralized word *kočky* 'cats' that appears in turns 17, 19, and 21 warrants a brief discussion. The participant encounters problems retrieving the lemma, which is mitigated by the interviewer's prompt in 18. However, a plural form is produced and in 21 corrected to singular. The distributional characteristics of the lemma in the ORAL are markedly different from *ryba*. The lemma *kočka* is singular dominant (65 % of occurrences in singular) and the citation form *kočka* is the most frequent (34.7 %). How can this error be explained under the usage-based model? The combined relative frequency of the form kočky (syncretic between plural nominative and accusative) is 26.3 % which is the second highest after *kočka* and would thus still be expected to be highly available for retrieval. Combined with possible structural priming effects due to the previous activation of the pattern [[R]-y] via *ryby*, it is highly likely that this form had a high potential of being "misretrieved".[25] This example is comparable to the cases discussed by Hatchard and Lieven (2019), who describe singularization and pluralization errors in connected speech of persons with aphasia, as discussed in the previous section.

The situation is slightly more complex in the case of the gender mismatch tokens. Let us start with the past tense verb forms. These are historically perfect participles that agree with the subject both in gender and number. While both singular and plural is marked for gender in standard written

25 Structural priming describes the situation where the use of a linguistic structure facilitates the use or processing of a similar structure due to "lingering" activation of the *prime* structure (e.g. Pickering & Ferreira 2008).

Czech, the gender distinction in plural is effectively leveled in the spoken language. It should also be noted that the distinction between animate and inanimate masculine nouns is not expressed in these forms, as shown in Table 5.

	singular		plural (written)	plural (spoken)
máma ‘mom’	zmizel-a	mámy	zmizel-y	zmɪzel-ɪ
táta ‘dad’	zmizel	tátové	zmizel-i	
kurf ‘briefcase’	zmizel	kufry	zmizel-y	
kotě ‘kitten’	zmizel-o	koťata	zmizel-a	

Table 5: Distribution of plural past tense forms with regard to gender in written and spoken Czech; on the example of zmizet 'vanish'.

When we look at the general frequencies of the singular masculine (animate and inanimate), feminine, and neuter forms of the participle, we see that the masculine forms are slightly more frequent. However, this does not necessarily explain the misgendering, because as can be expected, individual verbs have different preferences, as evidenced by Table 6, which shows the overall frequency for all verbs in a corpus of written Czech (Křen et al. 2020) (henceforth SYN2020) and relative frequencies of the three forms for the verbs *vidět* ‘see’, *vytáhnout* ‘pull out’, and *sežrat* ‘devour, eat’ that appear in turn 21.

verb lemma	gender	frequency pmw	relative frequency
all lemmas	F	17532.8	0.388
	MI	19421.8	0.429
	N	8276.0	0.183
vidět ‘see’	F	396.2	0.516
	MI	369.4	0.481
	N	1.7	0.002
vytáhnout ‘pull out’	F	19.2	0.395
	MI	27.9	0.575
	N	1.5	0.031
sežrat ‘devour’	F	8.2	0.321
	MI	15.8	0.620
	N	1.5	0.058

Table 6: Distributions of singular past participle forms accoring to gender for all lemmas and for selected verbs.

While *vytáhnout* and *sežrat* have more frequent masculines, both the feminine and the masculine are more or less equally frequent for *vidět*, all of which by itself shows that these word forms may enter in competition. However, the situation here is a little more complex than in case of the pluralized noun *ryba*, since the verb agrees in gender with the subject. In order to account for this fact, I sampled 200 most frequent nouns of the feminine and masculine (animate and inanimate together) gender from SYN2020 tagged as subjects which cooccur with a past participle and annotated these for animacy.[26] As shown in Figure 4, the probability of cooccurrence of a masculine participle with an animate subject is higher.

26 The queries used to obtain the data were [tag="NNFS1.*"&afun="Sb"&p_tag="V.FS....R.*"] and [tag="NN[MI]S1.*"&afun="Sb"&p_tag="V.[MI]S....R.*"] for feminine and masculine subjects respectively. While there are clear differences between the structure of written and spoken corpora, it should be noticed that both sources of data are representative of the linguistic experiences of Czech speakers to some extent. Also, a quick query of the spoken corpus, sampling 200 most frequent feminine and masculine nouns in singular nominative show that the difference in animacy between feminines and masculines is somewhat lower, but still holds (47.40 % animate masculines vs. 38.34 % animate feminines).

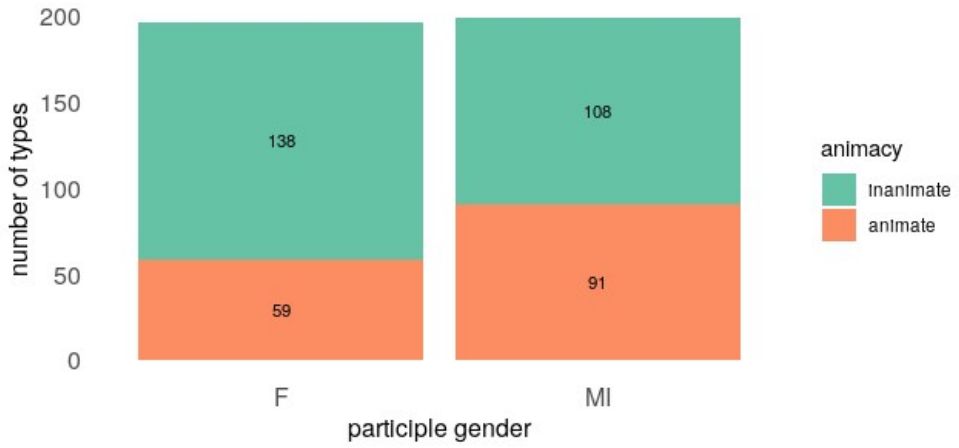

Figure 4: Distribution of animate and inanimate masculine and feminine subejcts of singular past tense verbs in syn2020. Numbers on columns show type frequencies.

Moreover, the probability of a feminine noun's denoting an animate referent is relatively low. Only 11.23 % of the 1630 feminine lemmas with a frequency greater than 10 pmw in the SYN2020 and 16.48 % out of 807 feminine lemmas in the same frequency range are animate. One more variable can be factored into this account, viz. the fact that *kočka*, the subject of the verbs in question, is an animal. Local and culturally important animals typically have lexemes denoting female (feminine), male (masculine animate), and young (neuter), e.g. *slepice* 'hen-F', *kohout* 'rooster-M', and *kuře* 'chick-N'. Gender is assigned on purely formal grounds for other animal lexemes, e.g. *kosatka* 'killer whale-F' vs. delfín 'dolphin-M'. I assembled a list of animal lemmas and annotated these for gender. One can conclude that the majority of animal lemmas is masculine (566 vs 415 feminine lemmas).[27] 786 lemmas from the list are attested in SYN2020 out of which 455 is masculine and 306 is feminine.[28]

These facts can also be used in the explanation of the gender mismatch in the demonstrative pronoun (*toho-M.ACC* instead of the target *tu-F.ACC*). The frequency distributions of the whole class of pronominal demonstrative modifiers, which in Czech agree in gender, number, and case with the head noun, does not point at any meaningful patterns with regards to the difference between genders. The construction *ta N-F* has on the whole higher type and token frequency than the construction *ten N-M*, as well as the relative frequency of the lemmas that appear in this construction relative to the total number of lemmas of the given gender. The form *tu* is considerably more frequent than *toho* both when case and gender are taken into account and when the word forms as such are compared.[29] There are also no meaningful patterns of association between these demonstratives and spe-

27 The list was created by combining the Czech Wikipedia entries under the sections *Seznamy živočichů v Česku* (https://cs.wikipedia.org/wiki/Kategorie:Seznamy_%C5%BEivo%C4%8Dich%C5%AF_v_%C4%8Cesku) and *Plemena zvířat z Česka* (https://cs.wikipedia.org/wiki/Kategorie:Plemena_zv%C3%AD%C5%99at_z_%C4%8Ceska) with the list on the website http://www.zoologie.frasma.cz/abecedni%20seznam.html (last access 12-05-2019).

28 It should be noted that the number of masculine and feminine animal lemmas with a frequency above 10 pmw is comparable.

29 Based on the spoken corpus, *tu* has a frequency 1431.5 pmw and *toho* 50.9 pmw in the masculine animate accusative reading, cumulative frequency of the word form is 876.8 pmw; *toho* is syncretic and is also the form used for masculine (both animate and inanimate) and neuter singular genitive.

cific lemmas.[30] This is probably not that surprising, given the fact that the "short" forms of demonstratives *ten*, *ta*, *to* behave very much like definite articles in modern Czech (cf. Zíková 2018). Differences in distributions thus can be said to reflect differences in the distributions of both whole genders and individual lemmas across cases. This in turn is connected to the functions of the cases and the semantics of the nouns. For instance, the low number of uses of *toho* in the accusative is not very surprising, given that this word form is only used with masculine animate nouns and that animacy is strongly correlated with agentivity and subjecthood. Similarly, the high frequency of *tu* is caused by the fact that the feminine gender has the highest type frequency both in the spoken and written corpora and the proportion of animate referents is relatively low, as noted above. However, one must also take into account that both the use of *tu* or *toho* in an accusative context, such as following a transitive verb, are reliable cues that the following noun is the object of the verb and, on top of that, *toho* is a very reliable cue that the following noun refers to an animate being. Furthermore, the word form *toho* has a partial formal overlap with other adnominal modifiers agreeing with the head noun, e.g. *mojeho/mého* 'my-M.GEN/ACC' or *malého* 'little-M.GEN/ACC', whereas the form *tu* does not (cf. *moji/mou* 'my-F.ACC' or malou 'little-F.ACC').

Based on these facts, it is thus possible to propose a simple model of the situation which is firmly grounded in the usage-based approach. The speaker with aphasia is telling a short story with the support of pictures. While producing the target sentence *Kočka viděla a vytáhla tu rybu*, the speaker relies on the fact that both the agent and patient referent are animate. Under the assumption of reduced processing resources in aphasia, it may be less demanding for the speaker to use the masculine forms. These are more strongly associated with animacy than the feminine gender, which in turn may reduce the uncertainty of the speaker and their interlocutor even though the produced utterance is agrammatical. I interpret this situation as an alleviating or compensatory mechanism which is grounded in the linguistic experience of the speaker. This may be further enforced by the existence of the abstract schemas [R-a viděl] and [toho R-u] represented e.g. by *táta viděl* 'dad saw' and *toho tátu* 'the dad-ACC' which may be expected to be attested in the speaker's linguistic experience, as the frequency of the masculine nouns of the type *táta* is not insignificant.[31]

In conclusion, I hope to have provided a detailed overview of the way language data can be described and analyzed within the usage-based framework. By accounting for various measures of frequency as well as the context of occurrence of the analyzed words in the discourse, it is possible to explain the two types of paragrammatic errors using a single general model that combines frequency effects with priming and animacy effects, demonstrating some of the advantages of this approach.

30 Possible such associations were assessed by using data from both ORAL and SYN2020. I queried both corpora so as to obtain the frequency data for the uses of demonstrative pronouns used as prenominal modifiers. The measures of *attraction* and *reliance* (Schmid & Küchenhoff 2013) were subsequently obtained for this data.

31 In fact, there are four such lemmas among the 20 most frequent masculine animate nouns in oral: *táta* 'dad', *děda* 'grandpa', *brácha* 'brother', *strejda* 'uncle'.

3 Building a corpus of Czech aphasic speech

3.1 Corpus assembly

One of the aims and outcomes of the present work was the assembly of a corpus of connected speech of Czech speakers with aphasia that would be available to students, researchers, and SLTs and would thus help to facilitate and entice further interest and research of aphasia in Czech which has been limited to a relatively small number of studies (cf. section 2.3). A further objective was to create a simple and user-friendly elicitation and transcription protocol to enable SLTs to contribute own data and help expand the corpus.

Collection of aphasic data is notoriously complicated and posits specific demands with regard to research ethics. The target population is relatively small, highly heterogeneous both in terms of intra- and interindividual variation, difficult to approach, and sensitive due to reduced levels of language comprehension. Data collection is thus highly demanding both in temporal and personal resources. A possible way to partially tackle these problems is the construction and publication of corpora of aphasic speech, i.e. relatively large, general purpose databases that can provide researchers from different fields and theoretical backgrounds with samples of speech characteristic of different aphasia types. Such data can be used directly to produce analyses or it can serve to generate hypotheses and predictions that may be tested in specifically designed protocols.

Currently, there are only few corpora of aphasic speech worldwide, mostly for the English language. AphasiaBank (MacWhinney et al. 2011) is by far the largest, most important and ambitious of existing corpora of aphasic speech. The corpus contains the recordings of 258 individuals with aphasia and is built on the CHILDES engine (MacWhinney 2000). The samples follow a single, well-defined protocol and are transcribed according to the CHAT rules, with special tags for aphasic errors. Each participant performs a battery of tasks which target areas typically problematic in aphasia, such as naming or repetition. There are also samples of free connected speech. The database was originally developed for English, however other languages such as Spanish or Mandarin Chinese have been included. The Cambridge Cookie Theft Corpus (Williams et al. 2010) contains speech samples of speakers with post-stroke aphasia but also other etiologies, elicited using the Cookie Theft picture, a stimulus picture for testing narrative discourse production in individuals with aphasia, which is a part of the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan 1972b). Samples from neurotypical speakers are also included for comparison. A Corpus of Dutch Aphasic Speech modeled after spoken corpora of unimpaired spontaneous speech has also been reported to be under development (Westerhout 2005; Westerhout & Monachesi 2007).[32] A corpus of Greek aphasic speech is also being developed (Varlokosta et al. 2016). The GREECAD corpus contains speech samples which are morphologically tagged and annotated for narrative structure and related discourse-level phenomena. Two relevant corpora of Slavic languages have also been developed. CliPS (Khudyakova et al. 2016; Bergelson & Khudyakova 2020) is a corpus of 39 Russian speakers with aphasia and five speakers with right hemisphere damage. Both audio and video

32 Note that both the Cambridge Cookie Theft Corpus and the Corpus of Dutch Aphasic Speech were reported in the literature to be in development, but neither project has been, to the best of my knowledge, published as of the time of writing.

recordings are included in the lemmatized and morphologically tagged corpus. The lemma tier also includes English translations. CroDA (Kraljević, Hržica & Lice 2017) is a corpus of Croatian aphasic speech that was developed in accordance with the AphasiaBank standards and includes the connected speech of 20 Croatian speakers with fluent and non-fluent aphasia.

3.2 Corpus design

The structure and design of the present Corpus of Czech Aphasic Speech has been informed by the design of available corpora of aphasic speech as well as the design of both a general-purpose corpus of spoken Czech (ORAL) and specialized corpora of Czech that form part of the Czech national corpora maintained by the Institute of the Czech National Corpus, such as the corpus of Czech-as-second-language (Šebesta et al. 2014).

To provide prospective users with examples of different genres, I decided to collect samples of descriptive, narrative, procedural, and conversational discourse.[33] The conversational discourse was based on three broad topics: participants' personal or professional life, hobbies, and speech language therapy. The descriptive discourse was elicited using four pictures with two levels of difficulty. The pictures were taken from the works of Josef Lada whose distinctive style was assumed to be relatively easy to understand for the participants. Two black-and-white drawings were selected in the low difficulty condition that depict two people in a static position with no or very limited background. The high difficulty pictures were two colored drawings depicting rich scenes with many participants and events happening in parallel. Both of the drawings depict scenes from life in the countryside in the early 20th century (a hog-killing scene and a scene from a village pub with dancing and music). One additional picture drawn by me was used during some of the sessions. This picture depicts a lady making a threatening gesture at a group of children after they kicked a football into her flowerbed. The picture was designed so that it demands participants to employ causal inference in order to fully understand and describe the picture. Three tasks were used to elicit narrative discourse. The lion cage scene, approximately three minutes long, was clipped from the Charlie Chaplin 1928 movie *The Circus* and used in a story retelling task. Participants watched the video on an eight inch tablet computer and were instructed to retell the story with as much detail as they would remember. The second task for narrative discourse elicitation used two three-frame comic strips with low and high difficulty prepared by me. The participants were instructed to create a short story based on the pictures. The low difficulty item used a simple background that did not change across pictures and the pictures in the strip showed directly following scenes in the story. The more difficult stimulus, on the other hand, depicts a more complicated story with two miniature story lines and a complex and changing background. There are also gaps in the story that require causal inference. Thirdly, I asked participants to recollect the events of the Velvet revolution in 1989 and the influence of the events on their professional or personal lives. One additional task was used during some of the sessions (pa1, pa2, pa3, aa1, and aa3). Following the AphasiaBank protocol (MacWhinney et al. 2011), I asked participants to tell the story of the Little Red Riding Hood or, alternatively, the Nativity story. However, some of the speakers responded negatively to this task and I decided to exclude it from the remaining sessions. Finally, to elicit procedural discourse, I asked participants to describe how they would prepare a simple meal of their choice or a cup of tea or cof-

33 This decision was motivated by the fact that discourse type has been shown to influence the performance of speakers with aphasia (e.g. Olness 2006).

fee. This task was later replaced with a map task to collect a more controlled and comparable production. One additional task was initially included (only in sessions pa1 and pa2) and aimed at participants' (meta)pragmatic skills. The two simpler drawings used to elicit descriptive discourse are accompanied with short texts jokingly explaining the depicted scene. After finishing the task, I showed the pictures to the participants again, read the texts, and asked the participants to explain the jokes to me. However, these failed to elicit expected responses and were excluded from the succeeding sessions.

Not all of these tasks were included in every session with every participant. The structure of individual sessions was adjusted to the severity of participants' aphasia, their reaction to the tasks, and time available for the session. The tasks that were selected for the protocol for future sessions are the following:

- conversation (hobbies, therapy)
- retelling of the lion cage clip
- both comic strips
- two pictures (Hat behind rails and Hog-killing)
- map task

The sessions proceeded as follows. The interviews were recorded in a campus office, in SLT offices, or, in one case, in the participant's home. Participants' SLT or significant other was present during some of the sessions (pa1, pa2, pa3, aa3, ba3, ba4). The sessions were recorded using a Sony DCR-SR70 video camera placed such that participants' upper body was visible. All interviews were administered by me. Prior to recording, I described the goals of the project, the structure of the session and the tasks and I explained to the participants how the data will be processed, presented, and stored. I also informed them about personal demographic and anamnestic data to be included in the corpus. This was done using an informed consent form.

The informed consent form was prepared with the aim to include the participants in the process and to make the form as accessible as possible. Following the guidelines from the literature (Kagan & Kimelman 1995; Palmer & Paterson 2011; Jayes & Palmer 2014), I used a larger font (14p.), technical language was kept to the minimum, and information was presented in bullet points in simple sentences (cf. Appendix 1). To further facilitate comprehension, I explained the information in the form also orally, checking for comprehension and answering any questions. Since there had not existed a research ethics committee at the Faculty of Arts at the time of data collection, the project could not receive ethical approval by the University, however, the project was reviewed and approved by attending SLTs before I approached potential participants or, in the case of the sessions recorded at the University Hospital Brno, by the hospital ethics committee.

After obtaining informed consent, the session started with the collection of participants' personal information (age, education, occupational area, place of birth, region of parents' birth, brief language biographies). Clinical information (etiology, time post onset, lesion localization, assessment after onset and at time of recording) was obtained from participants' SLTs. The order of the component

tasks was not fixed but rather adjusted to the situation. Because the aim of the project was to elicit as much production as possible, I interfered during the tasks when participants encountered difficulties to provide cues during word finding problems, to ask questions that would direct the participant toward the desired output, and to work collaboratively on repairs when necessary. In this way, a total of seven hours and 56 minutes was recorded with individual sessions ranging in length between 27 and 60 minutes. The data was collected between May and September, 2015.

Prior to data collection phase, I tested the tasks and session structure in one pilot session with one individual with aphasia. This recording is not part of the corpus. The participant, who has fluent chronic aphasia with mild-to-moderate comprehension deficits, reacted well during the pilot session. However, as described above, adjustments were made during the data collection phase in reaction to participants' behavior and specific needs during individual sessions or based on reactions and comments of three or more participants.

3.3 Participants

Eleven native speakers of Czech with chronic aphasia participated in the interviews. Participants were approached and recruited in three different places: through *Klub afasie*, a non-profit, outpatient therapy group based in Prague, at the Písek Hospital, and the University Hospital Brno. No specific inclusion or exclusion criteria were formulated beforehand, as an opportunity sample seemed as the only feasible way to recruit participants, given the specifics of the population and the state of the field in Czechia.[34] The project and its goals were described to the SLTs of the three institutions that agreed to participate in the project and the SLTs provided contact with clients with suitable language profiles and expected willingness to participate. Both fluent and nonfluent patients were included in order to obtain a range of patterns characteristic of different aphasia types. All participants had normal or corrected-to-normal vision and hearing. General characteristics of the participants are given in Table 7. To provide comparison with neurotypical speakers, three individuals with no recorded history of linguistic or other cognitive impairment who matched the demographics of the participants with aphasia were additionally recorded, using the three tasks that were selected for the analyses presented in section 4.

participant	gender	age	education level	m/y of onset	handedness	etiology	lesion site	assessment post onset	assessment at time of recording
aa1	M	69	S	n/a/07 (prev. n/a/98)	R	TBI	temporalne vlevo (prev predni jadra thalamu)	Global	Anomic

34 More SLTs had been approached, however, some had not responded at all, while others had shown little understanding toward the project.

aa2	M	74	S	4/14	R	iCVA	povodi ACM lsin	TMA	TMA
aa3	M	69	T	12/03	R	TBI	(subdur hemato ma) FTPO lsin fol-lowed by (sub-arach hem-orrh) T lsin	N/A	Conduc-tion
aa4	M	63	T	03/09	R	iCVA (ACM sin ACI sin)	post-malat-icka cysta capsula interna sin a basipari-etalne starsi is-chemie T-O-P	TMA	TMA
ba1	F	41	S	11/13	R	iCVA	L (P-O vlevo??)	nonflu-ent	border-line flu-ent Broca's
ba2	F	63	S	3/04	R	iCVA (vasculi-tis)	T-P vlevo	global	Anomic
ba3	M	77	T	5/15	R	iCVA	N/A	Nonflu-ent	Anomic
ba4	M	57	T	12/02	R	TBI	SDH, ICH FTP vlevo	Nonflu-ent	Broca's
pa1	F	66	P	2/15	R	iCVA	frontálně vlevo v povodí ACM ko-rtikálně + sub-	N/A	Anomic

							kort. + ložiskop arietook-cipitálně kor-tikálně		
pa2	M	52	S	8/01	R	hemor-rhCVA	basal ganglia	N/A	TMA
pa3	M	84	T	3/15	R	iCVA	frontálně vlevo v povodí ACM ko-rtikálně + sub-kort.	N/A	Concuc-tion
ac1									
ac2									
ac3									

Table 7: Demographic and clinical information about participants

3.4 Processing of recordings

All recordings were exported as .MPEG files, audio tracks were extracted in the .wav format. Prior to data collection, I decided to only include the audio recordings in the corpus. The reason for this was mainly practical. After discussions with SLT and first approach of prospective participants, several expressed that they would not be comfortable if videos were to be published as part of the corpus. Because it was expected that recruitment of participants would be complicated given the population, I decided to collect video recordings to provide additional support during transcription, but to publish only the audio in the corpus.

The transcription was performed in ELAN (Wittenburg et al. 2006). Each recording was segmented into turns and these into c-units (Loban 1966), communication units which are derived from T-units (Hunt 1966) but do not presuppose structurally full sentences with predication, i.e. they vary in size from single word turns to multi-clause units with rich inner structure. c- or T-units are oftentimes encountered in brain damage research (e.g. Coelho et al. 2012) and are seen as methodologically convenient and analytically valuable units in analyses of written and spoken discourse (see also Hatchard 2021: chap. 2 for a discussion of different approaches to segmentation in clinical linguistic research). For the purposes of the transcription, a c-unit was defined as a main-clause-like predication with all dependent predications or, in cases of a missing verb, as semantically cohesive wholes. Semantic cohesion was preferred over strictly formal properties in more complicated seg-

mentation decisions. For example, false starts and clauses with reformulations resulting in a main clause with two or more predicates structurally, but one proposition functionally were treated as constituting a single segment. Silent pauses between segments were computed automatically using ELAN's built in functions.

Transcription rules were based on the system used for the transcription of recordings for corpora of spoken Czech (e.g. ORAL). This was in part motivated by the fact that the corpus is planned to be published as part of the family of corpora maintained by the Institute of the Czech National Corpus. However, several adjustments were made due to the specific nature of the data. Specifically, a more nuanced transcription of disfluencies was used. Specific symbols were used to transcribe laughter, tuts and sighs, syllabification was also marked, and silent pauses were categorized by length. All silent pauses exceeding 200 ms were marked. This threshold was selected in accordance with interactional research and studies of behavior in conversation (e.g. Campione & Véronis 2002; or Fors 2011). While these results were published for neurotypical speakers and generally longer pauses and slower rate of speech is expected in some aphasia types (e.g. Hatchard 2021 in her analysis of fluency in aphasia disregards pauses shorter than one second), it was decided to use this threshold to enable profiling of speakers based on pause duration distributions. Additional duration categories included 500 ms, 1 s, 3 s, 5 s, and 10 s. In accordance with transcription rules for ORAL, orthographic representation was used when the speech adheres to orthoepy. Colloquial or regional variants were transcribed using the spelling system of Czech and non-phonemic lengthening and phoneme elision or epenthesis were transcribed accordingly. Phonemic paraphasias (e.g. voicing or cluster reductions) and neologisms were transcribed to resemble the production of the participant as closely as possible using Czech orthography with some additional symbols or diacritics when necessary. Admittedly, the level of phonetic detail is relatively coarse, because the corpus is focused on higher levels of language structure and some of the recordings were made in suboptimal conditions.

The transcripts were subsequently tokenized and prepared for lemmatization and morphological tagging. This was performed with the MorphoDiTa tagger (Straková, Straka & Hajič 2014) using a dictionary that was used to tag ORAL.[35] Given the specific nature of the data, the tagging was not without problems. While some of the errors were caused by differences between transcription systems and were relatively easy to correct (semi)automatically, other errors required manual correction. Most of these included verbs or pronouns that were tagged as proper nouns (e.g. *šel* 'went-SG.M' miscategorized as a form of a hypothetical nominal lemma *Šel*). This may have been in part due to single word utterances that occur very frequently in the data. The whole corpus has the parameters summarized in Table 8.

participant group	c-units	positions	word tokens
administrator	4355	20437	16954
speakers with aphasia	4449	31422	23164
neurotypical speakers	279	2390	1897

Table 8: Summary of the size of the corpus; word tokens are fully formed words (including paraphasias)

35 I would like to thank David Lukeš from the Institute of the Czech National Corpus who helped me with this processing step.

A subset of the corpus that was mainly used for the analyses presented in following chapters was corrected manually (lemma and part of speech only, not the full morphological tags) and annotated for a host of other variables, which are discussed in the next section. The manual correction allowed for a comparison and estimation of accuracy of the tagger for this sample. Original and corrected tags and lemmas were compared for 7627 tokens in the subcorpus. It should be noted that I did not compare the whole 16-position tags used in the ICNC corpora, only the part-of-speech tags. POS tagging was accurate in 93.93 % and lemmatization in 92.22 % of tokens. There were a number of paraphasic tokens and tokens transcribed in a way that was opaque to the tagger. When these were excluded, the accuracy increased to 96.12 % and 95.08 % for POS and lemma respectively.

In case of paraphasic and neologistic tokens, I provided lemma and/or part-of-speech tags in all cases where a reconstruction of the target was possible. The forms **lemma* or *?lemma* were used for two decreasing levels of certainty to clearly mark reconstructed lemmas in the corpus. In a small number of cases, it was not possible to reconstruct the target lemma, but it was still possible to recover part of speech of the paraphasic token based on the paraphasic form and the context. This was mostly the case in tokens bearing clearly recognizable inflectional and/or derivational morphemes (e.g. the past participle marker *-l*). Examples 3.a-c show in sequence a token with reconstructed lemma and part of speech, with part of speech only, and a token where neither was recoverable.

3. Reconstruction of lemmas and parts of speech

    a. *votamřit* (aa3: 83) reconstructed as the verb otevřít ‘open’ based on formal similarity and the communicative context

    b. *nenečka* (ba4: 80) lemma not reconstructed due to unclear referent and lack of formal similarities to existing lemmas, reconstructed as a noun based on the ending *-(e)čka* typical of this word class

    c. *peram* (aa2: 249) not reconstructed due to the lack of context and formal similarities to existing words and morphological patterns

3.5 Structure of subcorpus annotation

A sample of the corpus data described in the previous section was selected for the analyses presented in this work. I describe the subcorpus and the annotation that was used for further analyses, then I characterize the subcorpus in terms of a suite variables used in the analysis of discourse production in section 4.2. The parts of the corpus that were used in the analyses include the retelling of the Lion cage scene with Charlie Chaplin (henceforth referred to as CHAPLIN), the description of the hog-killing picture (LADA), and the story creation based on the three picture strip with cat and cake (COMIC). These tasks were chosen as part of the core session material and present relatively long samples of directly comparable speech.[36] The reasons for the selection of the former two tasks was to include an example of both descriptive and narrative discourse. The story creation was included for two reasons. Firstly, participants were explicitly instructed to produce a narrative (*příběh* ‘story’,

36 This is in contrast both to conversational discourse (different topics spread across whole sessions) and procedural discourse (not all participants saw the same stimuli).

co se tam stalo 'what happened'),[37] but the static nature of the stimulus (as opposed to the video clip) resulted in the production of a descriptive text instead in some of the participants. Secondly, the task had proven to be quite difficult for the participants with aphasia, most likely because parts of the story were only implied and participants had to use causal inference to "fill in" the gaps, and this provides good grounds to analyze the language production in the context of a demanding task/ increased stress.

The subcorpus contains 1396 c-units produced by participants (i.e. excluding speech produced by the administrator) and 11366 tokens of which 7648 are fully formed words (including paraphasic words and neologism), 1027 c-units were produced by the administrator or another person present during the interview which corresponds to 5150 tokens out of which 3962 are fully formed words. In order to enable further analyses, this subcorpus was further processed and annotated for a suite of variables. A detailed description of this coding scheme follows.

The transcript, lemma, and morphological tag layers were complemented by error annotation used to mark and roughly classify aphasic errors. A simple coding scheme was used in which whole c-units (instead of individual tokens) were marked for errors. Errors were categorized as form, meaning, grammar, and discourse errors. This approach circumvents the question of tag placement in the case of errors spanning multiple words, omitted words, or errors concerning discourse coherence. On the other hand, the coding does not directly indicate where the error is located in the segment, which may be confusing in cases of multiple error occurrences. This also means that the coding is just a binary yes/no scheme with no indication as to the number of the given error type occurrences. However, given the limitations of time and personal resources, I believe that this system is clear and informative enough in most of the cases and erroneous tokens are identifiable easily.

A form error tag was used whenever a token deviated from the pronunciation expected in either the standard or a colloquial/regional variety of Czech. Such errors include phoneme alternation, deletion, or addition or a completely neologistic token. Thus, in 4 the form *stotole* is coded as form error, while *neim* is not, because it is a variant occurring frequently in spoken Czech.

4. Example of pahraphasic and colloquial production

    a. *stotole* (aa2: 382) is a paraphasic production of the target *stodole* 'barn-SG.LOC' with an inappropriate devoicing

    b. *neim* 'dunno' (pa2: 52) is a standard reduction of the word form nevím 'I don't know' used in spoken Czech

Meaning error tags were used in three cases. First, semantic paraphasias and circumlocutions were coded as such (cf. 5.a and 5.b respectively). Second, the tag was also used to mark clausal fragments where a participant encountered lexical retrieval problems and, unable to retrieve the word successfully, voluntarily leaves the segment unfinished. This is typically marked by very long silent pauses and repeated hesitations, as shown in 5.c. Finally and somewhat similar to the previous case,

37 Such instructions have been found to directly influence the nature of discourse production cf. Olness (2006) who found that participants produced more past tense and verbs that had a high degree of association with the storyline when explicitly instructed to use temporal sequencing.

segments with explicitly verbalized retrieval problems and unsuccessful target retrieval were marked as meaning errors (5.d).

5. Examples of lexical (meaning-driven) errors

    a. *hledal 3 zámek nebo H* ‘he was looking for 3 the lock or H’ (aa2: 29), the target was most likely *klika* ‘doorknob’ or *klíč* ‘key’, produced in the context of Chaplin locked in a cage semantic.paraphasia

    b. *takovej ten H 0.2 taky se na tom žehlí* ‘some such H 0.2 it’s also used for ironing’ (aa1: 118), the target was *prkno* ‘board’

    c. *a- ale 0.2 potom 0.5 no 1 <tut> H* ‘b- but 0.2 then 0.5 well 1 <tut> H’ (aa4: 78), produced while recounting how Chaplin was trying to bring a woman who fainted to

    d. *ale 0.2 já nevim jak se H jak se to řekne ...* ‘but 0.2 I dunno how to H how to say it ...’(aa3: 90), produced while trying to retrieve the verb *omdlít* ‘faint’

Grammar error tags were used for agrammatic and paragrammatic production, such as the omission of a preposition and an inflectional suffix in 6.a. and the substitution of the target reflexive pronoun form *si* with *se* 6.b.

6. Examples of grammatical errors

    a. *stojí tam 0.5 H tam 0.5 stůl* ‘stands there 0.5 H there 0.5 table’ (aa2: 164); this c-unit was produced while the participant was describing an acquarium that was on a table, based on the context, the expression *tam stůl* is taken here to have been produced instead of the target *na stole* ‘on the table’ and is analyzed as omission of the preposition as well as the inflectional morpheme

    b. *povídají se* ‘they are talking’ (ba2: 230), while this form exists in Czech in the anticausative construction, the target grammatical form here is *povídají si* with the dative form of the reflexive pronoun

Finally, the discourse error tag was used in two types of local incoherence. First, the tag marks c-units with unclear referential relations in which it is not possible to establish which entity the participant refers to. It should be noted, though, that this can in some cases be recovered from situational context as the picture stimuli were present in two out of three tasks. These cases may have been triggered by this as a relief mechanism. Second, segments with low information value (empty speech), typically using general all-purpose vocabulary were coded as discourse errors. 7.a. is an example of the former type, 7.b. illustrates the latter.

7. Examples of discourse level errors

    a. *to bylo dvaná- dvě hodiny / tam dávala* ‘it was twel- two o’clock / she was putting in’ (aa3: 220-221), the participant it beginning here to create a story and uses a pro-drop reference without having introduced any participants to the story previously

b. *tak tadydleten to byl tedy čárl- s- saplin 0.2 z tadytoho hlavního 0.2 jako jedinýho 0.5* ‘so this one was then Char- s- Saplin 0.2 from this there main 0.2 like only 0.5’ (pa3: 3), despite identifying the main character of the story the participant does not produce a contentful predication

The data was further processed prior to analysis. Lemmatization and morphological tagging for all tokens was checked and corrected manually, as already discussed, as were parts of the transcripts containing typing errors. Automatically created silent pauses between segments produced by the same speaker (intra-turn pauses) were integrated with corresponding preceding c-units. Additionally, there are 563 between-turn pauses that constitute a separate tier. Segments that constituted a single c-unit but were split into two during transcription due to an occurrence of a backchanneling segment of the interlocutor, such as 8, were merged into single segments which allowed for a more precise computation of c-unit length.

8. Merged c-unit that was transcribed in two separate segments due to an intervening interlocutor turn

   a. original transciption:

      aa1: 194: *a: 0.5 čuchá tady pes* ‘and 0.5 a dog is sniffing here’

      exp: mh

      aa1: 195: *k podlaze 0.2* ‘the ground 0.2’

   b. after merging:

      aa1: 194: a: 0.5 čuchá tady pes k podlaze 0.2 ‘and 0.5 a dog is sniffing the ground here 0.2’

      exp: mh

The data was coded for a range of variables that were entered into the analysis. These variables are derived from various measures of fluency used in quantitative analyses of discourse production as discussed in section 4.2. The variables are grouped into conceptually similar clusters.

- Disfluencies and repetitions: Hesitation sounds, tuts, silent pauses, and word fragments were categorized using specific part of speech tags and counted as disfluencies. Coughs and laughter were disregarded. Repetitions were also included under disfluencies. While it was possible to extract the number of immediate direct repetitions automatically from the data, repetitions spanning multiple words or repeated words separated by pauses or hesitations were counted manually.

- Fragmentary segments: All c-units that were fragmentary were marked as such. The analysis took a pragmatic functional approach to what constitutes a sentence fragment, rather than a formal structural one (cf., inter alii, Kynette & Kemper 1986; Walker, Roberts & Hedrick 1988; Kemper et al. 1989). This was motivated by the nature of spoken data and, mainly, by the aims of the analysis. Pragmatically felicitous, functional equivalents of full sentences

(e.g. single word answers to questions) were treated as characteristic of general spoken language, whereas abandoned c-units, single word utterances, or listings of nouns in the form of an enumeration lacking a predicate were taken as indicating language deficit.

- Length: In order to measure mean length of c-unit (MLU) and average number of disfluencies per c-unit, the following values were computed: total length in tokens, total number of fully formed word tokens (excluding paralinguistic backchanneling tokens), and total number of disfluencies (silent pauses, hesitations, fragments, repeated tokens). Paralinguistic tokens such as laughter, coughs, or sighs were ignored.

- Corpus frequency data: Lemma frequencies normalized per million words based on ORAL and a corpus of written Czech (Křen et al. 2015) (SYN2015) were obtained. Frequency of lemmas without corpus attestation was coded as zero and such lemmas were excluded from analyses of frequency.

- Grammatical tags: Simplified grammatical tags were derived, corrected, and adjusted from the morphological tags to differentiate between subgroups of items with different functional distributions within parts of speech, such that only the information relevant for the analyses was retained and certain items were recategorized, these tags were also designed to be "human readable". Gender, number, and case were tagged for nouns and adjectives. Numerals and prepositions were not further categorized. Pronouns were categorized as personal, reflexive, relative, interrogative, demonstrative, indefinite, possessive, and negative, pronouns introducing complement clauses were labeled accordingly. Verbs were coded as past (i.e. in past participle form) or nonpast, imperative, passive participle, or infinitive. The verb *být* 'to be' which is used as auxiliary in the expression of past tense, future tense of imperfective verbs, passive, and the conditional was marked as auxiliary or conditional in those cases and as past or nonpast when used lexically in a copula construction. This coding was performed semi-automatically. Conjunctions were grouped as coordinating or subordinating. Adverbials were coded as *adverbs proper* functionally and semantically or as adverbs introducing subordinate clauses.[38] The word *tak* 'so, such' is tagged as an adverbial in the CNC but has a range of different functions. Part-of-speech tag was changed to particle or conjunction when the word was used as a discourse marker or as a "resumptive" connective that appears in postponed main clauses (e.g. *Když tam přišel, tak se posadil.* and 'When he arrived, he sat down.'). Adverbials were annotated manually in this way.

- All nouns, adjectives and numerals, and all lexical verbs (i.e. excluding auxiliaries) and all lexical adverbs were coded as content words. In contrast to traditional categorizations, discourse markers, interjections, and back channeling devices were coded as interactional words. Remaining tokens were categorized as function words.

- Intervening segments: While the analysis is, naturally, focused on the production of the participants with aphasia, one of the variables used for the analysis is concerned with the production of the administrator. Specifically, administrator c-units were categorized as conver-

38 This category of adverbials functioning as complementizers is similar to the one used for pronouns. These lemmas form in the majority of cases a single category, but were tagged as either pronouns or adverbs by the CNC tagger. I did not amend this part-of-speech tagging, as this would not alter the results presented here in any way.

sational or as intervening prompts in situations when the examiner intervened to prompt the participant during retrieval problems or to participate in repairs. Backchanneling and clarifications related to the tasks were categorized as conversational units.

- Error coding: Apart from category specific variables, a generalized binary variable indicating presence of any type of error was used.

- Multi-clause units: Multi-clause segments were annotated automatically, using subordinator occurrence combined with finite verb count, manual checks were subsequently performed.

- Modifiers: To provide an additional measure of sentence complexity in the form of the expressed non-essential information, modifiers or backgrounding devices were counted. Modifiers were defined as adjectives, the pronouns with adjective morphology *nějaký* 'a/some' and *žádný* 'not any/none', possessive pronouns, numerals, prepositions, relativizers, adverbials, and subordinating conjunctions. The total number of modifiers was reduced by one for predicates that according to a valency dictionary of Czech (Lopatková et al. 2020) contain and adverbial as part of their argument structure. It should be noted that this is just a very crude estimate, as it does not, for instance, differentiate between adnominally and predicatively used adjectives and demonstrative pronouns are not included because they are frequently used in non-modifier function.

- Lexical diversity: Instead of using type token ratio (TTR) or moving average type token ratio (MATTR), I decided to use zTTR, a measure that modifies TTR by comparing the value to a distribution of types and tokens in reference texts of the same length proposed by Cvrček and Chlumská (2015). The spoken corpus was used as the reference corpus and only well-formed types and tokens were used to obtain the measure. The values were computed using an on-line app provided by the Institute of Czech National Corpus (Cvrček 2019).

- Reformulations: All c-units where participants reformulated either a single word, e.g. to repair a grammatical error, or the whole clause were coded as containing a reformulation.

- Story relevant vocabulary: To obtain an index of participants' ability to use specific vocabulary related to the stimuli, I obtained a list of core vocabulary items for the three tasks. The following procedure was used to obtain it. I created a questionnaire using Google Forms in which the three stimulus materials were included. Participants were instructed to describe the pictures and the video using eight to ten sentences. The questionnaire was distributed online and 19 responses were collected. Responses were lemmatized and part-of-speech tagged using MorphoDiTa. A list of stopwords was used to filter out general high frequency lemmas.[39] Lemmas that were used by more than a half of respondents, i.e. 10 or more, were labeled as story-relevant and the percentage of occurrence of these was used as a very crude index of informativeness. The lists of story relevant lemmas is provided in section 4.2.2.7.

Processed in this way, the data was prepared for further analyses that are presented in section 4.

39 The list was created by the Natural Language Processing Center at Masaryk University in Brno, available at https://nlp.fi.muni.cz/trac/research/wiki/cs/StoplistZakladnichTvaru.

4 Analyses

4.1 Introductory methodological remark

blah blah

4.2 Discourse production measures in aphasia research and clinical practice

Despite the fact that (interactional) discourse is the most natural form of language from the perspective of ecological validity, linguistic research of aphasia was for a long time focused on subword and single word and sentence level processing, while a rise of interest in discourse production and comprehension emerged later. This is somewhat similar to linguistics as a field in general. In their review of discourse analyses, Bryant and colleagues (2016) found that the number of discourse analyses had doubled in the period after 1995 compared to 1975-1995 and there was a further increase, particularly in treatment studies, in the late 2000s. In a follow up paper, Bryant et al. (2017) focused on the use of discourse analyses in clinical practice. Based on the responses of 123 SLTs from the English speaking world, they found that about 50 % of SLTs agreed that discourse analysis should be an important part of aphasia assessment in practice, however, only 30 % of the participant SLTs considered themselves competent to perform such analyses and only 25 % of SLTs reported the use of discourse analysis based on detailed transcriptions. Bryant et al. (2016) propose that clinicians may consider discourse analysis an important part of language profiling in aphasia but that there is a sense of imbalance between costs and benefits: the process of discourse elicitation, transcription and coding, analysis, and interpretation of results is time consuming and oftentimes requires special knowledge and the use of specialized tools. This issue has been addressed by several researchers who have proposed different solutions. For instance, Kim and colleagues (2019) suggest the use of core lexicon as a crude measure that does not require transcriptions. Casilio et al. (2019) show that auditory-perceptual ratings may be an efficient, reliable, and valid alternative that is strongly correlated to quantitative measures. Lind and collaborators (2009) devised a set of “user-friendly” linguistic indices that do not require extensive linguistic training and have the potential to differentiate between speakers with aphasia and neurotypical speakers.

The importance of discourse profiling in clinical practice is highlighted by recent findings regarding discourse processing in mild aphasia. While earlier results suggested that discourse structure (e.g. genre selection, basic storyline management in narrative discourse etc.) is preserved in mild and moderate aphasia (cf. Ulatowska, Allard & Chapman 1990’s review), more recent studies have demonstrated that persons with mild anomic aphasia (Andreetta, Cantagallo & Marini 2012), residual aphasia (Jaecks, Hielscher-Fastabend & Stenneken 2012), or individuals who score above the cutoff score for aphasia diagnosis in an assessment test (Fromm et al. 2017) encounter problems in discourse production and that their discourse can be characterized by a decrease in MLU, informativeness, coherence and cohesion, lexical diversity, open class word frequency, or an increase in the use of pronouns and formulaic expressions, compared to neurotypical speakers.

This area of aphasia research is characterized by a great variability of approaches and methods. Bryant and colleagues (2016) report over 500 measures of discourse productivity in their survey of 165 papers. Similarly, Linnik et al. (2016) conclude that this variability in used measures is reflected by varying definitions and scoring methodologies which may lead to lower reliability of the measures. Pritchard and collaborators (2017; 2018) focused on the psychometric properties of discourse production measures report that there is only a limited amount of psychometric information available in the literature.

While most of the analyses of discourse in aphasia have been quantitative (using text based measures or expert ratings), some authors have argued for the merits of Conversation analytical methods, particularly the organization of turn taking, adjacency pair structure and the management of repairs (Boles 1998; Wilkinson et al. 1998; Damico, Oelschlaeger & Simmons-Mackie 1999; Perkins, Crisp & Walshaw 1999). Quantitative analyses have focused on micro- and macrostructural characteristics of discourse (see Armstrong 2000; Prins & Bastiaanse 2004; or Linnik, Bastiaanse & Höhle 2016 for reviews). Microstructure analyses use a suite of measures that are based on clause- and word-level phenomena such as mean length of utterance, well-formedness of clauses, lexical diversity, or semantic and argument structural characteristics of verbs, i.e. features that revolve mostly around the concepts of fluency and grammaticality. Macrostructure has to do with the organization of discourse and derived measures are typically based on informativeness of the text, cohesion and coherence, or thematic components of discourse.

Recent analyses have increasingly focused on combining these two levels of structure in multilevel analyses. In a representative example, Marini et al. (2011) analyzed samples of 300-400 words combined microstructural measures (MLU, proportion of paraphasias, paragrammatic errors, phonological errors, omissions of grammatical morphemes, and grammatical clauses) with macrostructural indices (proportion of cohesive errors, local and global coherence errors, appropriate lexical information units, and thematic units) to describe the production of two individuals with aphasia and showed that these combined measures can reveal problems that are not detected by standard general assessment tools. With regards to cross-linguistic variation, Prins and Bastiaanse (2004) argue that available results suggest that the measures in use have cross-linguistic validity.

Widely used linguistic microstructural measures include:

- MLU
- number of tokens
- number of types
- TTR (and its variations such as MATTR; whole texts or nouns and/or verbs specifically)
- number of errors of different types
- number of embedded/subordinate clauses
- semantic verb types (typically based on Halliday’s categorization (Halliday & Matthiessen 2014: chap. 5)

- number of predicate arguments
- propositional density (typically implementations of Brown et al. (2008)'s automated version, cf. MacWhinney et al. (2010))

The fact that there are many different measures in use has led some researchers to investigate whether some of these variables may cluster together in terms of underlying processes responsible for those and to assess the relative importance and contribution of those variables. In an early example of this approach, Wagenaar and colleagues (1975) analyzed spontaneous speech data from 74 Dutch speakers with aphasia by using 30 different measures that were subjected to factor analysis and identified six clusters of variables with fluency being the primary dimension. They also concluded that binary fluency classification can be made based on two variables (rate of speech and MLU). Glosser and Deser (1991) submitted the performance of 44 participants (non-brain-damaged, closed head injury, probable Alzheimer's disease, and fluent aphasia) on eight variables related to syntax, lexical errors, cohesion, and coherence to a factor analysis and identified three factors with high loadings: coherence, lexical errors, and syntactic complexity. Rochon et al. (2000) performed principal component analysis on nine of the measures used in Quantitative Production Analysis (QPA). QPA is a widely used coding scheme designed by Saffran and colleagues (Saffran, Berndt & Schwartz 1989; Berndt, Schwarz & Saffran 2001) to characterize the connected speech of English speakers with agrammatic aphasia which is focused on clause structure and grammar. Rochon and colleagues identified two factors related to the production of propositional utterances and free grammatical morphemes. They subsequently used a cluster analysis to differentiate within a group of 37 speakers with Broca's aphasia. They identified four clusters of performance that were directly related to the two factors. In a recent analysis, Fromm and collaborators (2021) used the data of 306 speakers from AphasiaBank comprising transcripts of different discourse types and detailed demographic and clinical information. They used almost 50 discourse variables as well as participant data in a cluster analysis to identify groups of patients with similar performance patterns. Importantly, aphasia type using the standard classification system was not part of the variable list. They identified seven clusters in the data and used cluster membership as outcome variable in a random forest analysis with the aim to identify the variables that contributed the most to cluster assignment. This procedure identified the total number of tokens produced in free speech and number of closed class tokens produced in Cinderella retelling as the most important predictors. Fromm and colleagues link these variables to productivity and grammaticality respectively. An additional contributing variable was retracing (repairs and reformulations).

Following up on the literature, I present a characteristic of Czech discourse in aphasia in this section. To the best of my knowledge, this is the first description of connected speech in aphasia for Czech.[40] The measures and dimensions of description I used are focused on microstructure and will be presented in three categories that are labeled "fluency", "productivity", and "well-formedness". The measures are summarized in Table 9.

40 A discourse analysis of the production of Slovak speakers with Broca's and anomic aphasia was recently published by Kevická and collaborators (2020). They report that speakers with anomic aphasia produced less units and more semantic errors than both speakers with Broca's aphasia and neurotypical participants and that there was an increase in the number of units and words produced in Broca's aphasia in comparison to the other groups. Both aphasia types were characterized by fewer thematic units produced but use of cohesive ties was comparable to the neurotypical group.

fluency	mean length of c-unit
	mean number of disfluency clusters per word
	mean number of fluent trigrams
productivity	zTTR
	mean number of backgrounded and modifying elements
	proportion of multi-clause c-units
	mean number of verbs per c-unit
	proportion of high frequency content lemmas (> 100 pmw in ORAL)
	mean frequency of content lemmas
	proportion of keywords used
	proportion of content and function words
	mean proportion of nouns to total number of nouns and verbs in c-unit
well-formedness	proportion of clausal fragments
	proportion of experimenter interventions
	proportion of c-units with any errors
	proportion of c-units with repetitions and reformulations

Table 9: Measures of discourse production used in the analysis.

Participants were grouped into three groups according to fluency (fluent, nonfluent, neurotypical), based on SLT assessments. These groups were compared using the non-parametric Kruskal-Wallis test followed by post hoc pairwise multiple comparison using the Dunn's test with Holm correction for multiple comparisons. The results for the measures of fluency, productivity, and well-formedness are discussed below with focus on group as well as individual differences.

4.2.1 Fluency

The summary of individual fluency measures for each participant is provided for reference in Table 10. Particular measures and individual differences are discussed below.

partic-ipant	group	number of c-units	token fq (in-cluding repeti-tions)	MLU mean (sd); median (MAD)	fluent trigrams mean (sd); median (MAD)	disfluent chunks / word
aa1	fluent	103	790	7.515 (5.886); 5 (2.965)	3.427 (3.642); 2 (2.965)	0.178
aa3	fluent	168	957	5.381 (4.003); 4 (2.965)	1.345 (1.883); 1 (1.483)	0.352
ba2	fluent	122	663	5.254 (3.403); 5 (2.965)	1.189 (1.428); 1 (1.483)	0.485
ba3	fluent	114	759	5.991 (4.97); 5 (2.965)	1.746 (2.445); 1 (1.483)	0.346
pa1	fluent	87	488	5.414 (4.538); 4 (2.965)	1.713 (2.778); 1 (1.483)	0.236
pa3	fluent	106	806	7.057 (5.715); 5.5 (4.448)	2.311 (3.322); 1 (1.483)	0.270
aa2	nflu-ent	159	508	2.83 (2.087); 2 (1.483)	0.208 (0.585); 0	0.702
aa4	nflu-ent	77	225	2.714 (1.999); 2 (1.483)	0.286 (0.604); 0	0.560
ba1	nflu-ent	41	157	3.78 (2.859); 3 (2.965)	0.463 (0.869); 0	0.594
ba4	nflu-ent	89	178	2 (1.492); 1 (1.483)	0; 0	0.691
pa2	nflu-ent	51	221	4.235 (2.371); 4 (2.965)	1.196 (1.327); 1 (1.483)	0.241
ac1	nbd	108	915	8.065	4.167 (3.964); 3	0.169

				(5.488); 7 (5.93)	(4.448)	
ac2	nbd	108	621	5.611 (5.372); 4 (4.448)	2.287 (2.852); 2 (2.965)	0.239
ac3	nbd	63	360	5.571 (2.832); 5 (2.965)	2.778 (2.642); 2 (2.965)	0.154

Table 10: Summary of the fluency measures for individual participants.

Starting with the three measures of fluency, the comparison of the three groups is summarized in Table 11, where the Kruskal-Wallis test reached significance at alpha level 0.05, the results of the post-hoc pairwise comparison is included. We see that all of the used measures detect group differences. MLU measured in well-formed words with excluded repetitions differentiates between the neurotypical speakers and the non-fluent group on the one hand and between fluent and non-fluent speakers on the other. No significant difference was found between neurotypical speakers and participants with fluent aphasia. The remaining two measures (mean number of fluent trigrams per c-unit, mean number of disfluent chunks per words) differ between the neurotypical and non-fluent participants. The performance of individual participants is discussed below.

variable	Kruskal-Wallis (n = 14, df = 2)	pairwise comparison
MLU	$\chi^2 = 9.257$; $p = 0.0098$; $\eta^2 = 0.66$	fluent x non-fluent (z = -2.566; p = 0.0265); nbd x non-fluent (-2.619; p = 0.0265)
mean fluent trigrams	$\chi^2 = 0.363$; $p = 0.0093$; $\eta^2 = 0.669$	nbd x non-fluent (z = -2.88; p = 0.004)
disfluency per word	$\chi^2 = 8.069$; $p = 0.0177$; $\eta^2 = 0.552$	nbd x non-fluent (z = 2.793; p = 0.0053)

Table 11: Group comparisons of the measures of fluency.

4.2.1.1 MLU

Figure 5 shows the distribution of c-unit lengths in fully-formed words with excluded repetitions. The group pattern is clearly visible here with a typical c-unit in the non-fluent group containing less than five words. We see that the majority of c-units produced by ba4 were single or two word utterances with the few longer segments having mostly the form of listings of nouns. On the other hand, participants ba1 and pa2 in the non-fluent group produced on average longer c-units with median values of 3 and 4 words respectively. When we look at the fluent group and the neurotypical speak-

ers we see two slightly different profiles represented by aa1, pa3, and ac1 on the one hand and aa3, pa1, ac2, and ac3 on the other with the former group characterized by a greater number of longer c-units.

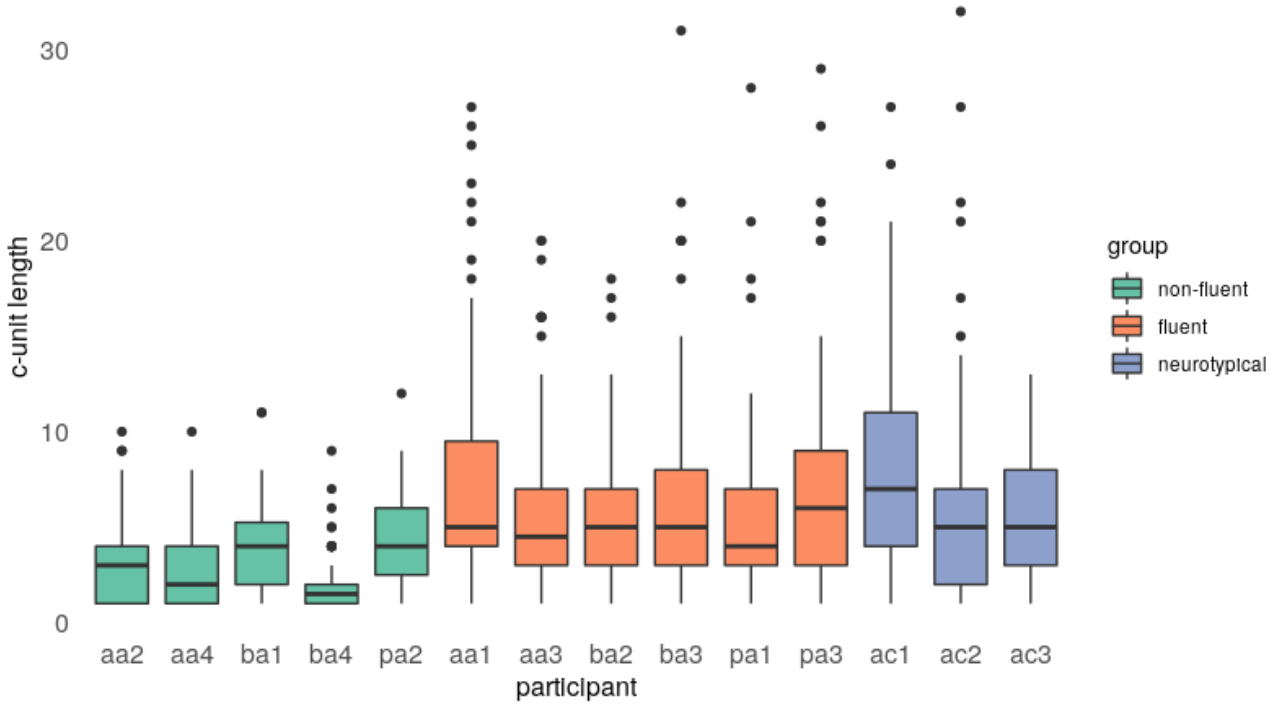

Figure 5: Boxplots showing the distribution of c-unit lengths for individual participants.

4.2.1.2 Fluent trigrams

The distribution of fluent trigrams across c-units is shown in Figure 6 (only the c-units with a length of 3 or more words are included, participant ba4 is excluded since he did not produce any fluent trigrams). It is important to keep in mind here that, by definition, longer segments have more trigrams. Participants with a median length of c-unit equal to 2 will naturally have produced very few trigrams on the whole. This is exactly what we see in participants aa2 and aa4. Participant ba1 who produced longer c-units also patterns here with aa2 and aa4, Similarly, when we focus on the fluent group of participants, we see that pa3 produced fewer fluent trigrams than aa1 even though they share a similar profile with regards to MLU. pa3 patterns together with ba3, while aa3 and ba2 share both the MLU as well as the trigram profile. Interestingly, participants pa1 and pa2 pattern together here which is driven by relatively by their comparable MLUs (5.4 and 4.2 respectively) and a similar number of disfluent chunks per word (0.236 and 0.241 respectively).

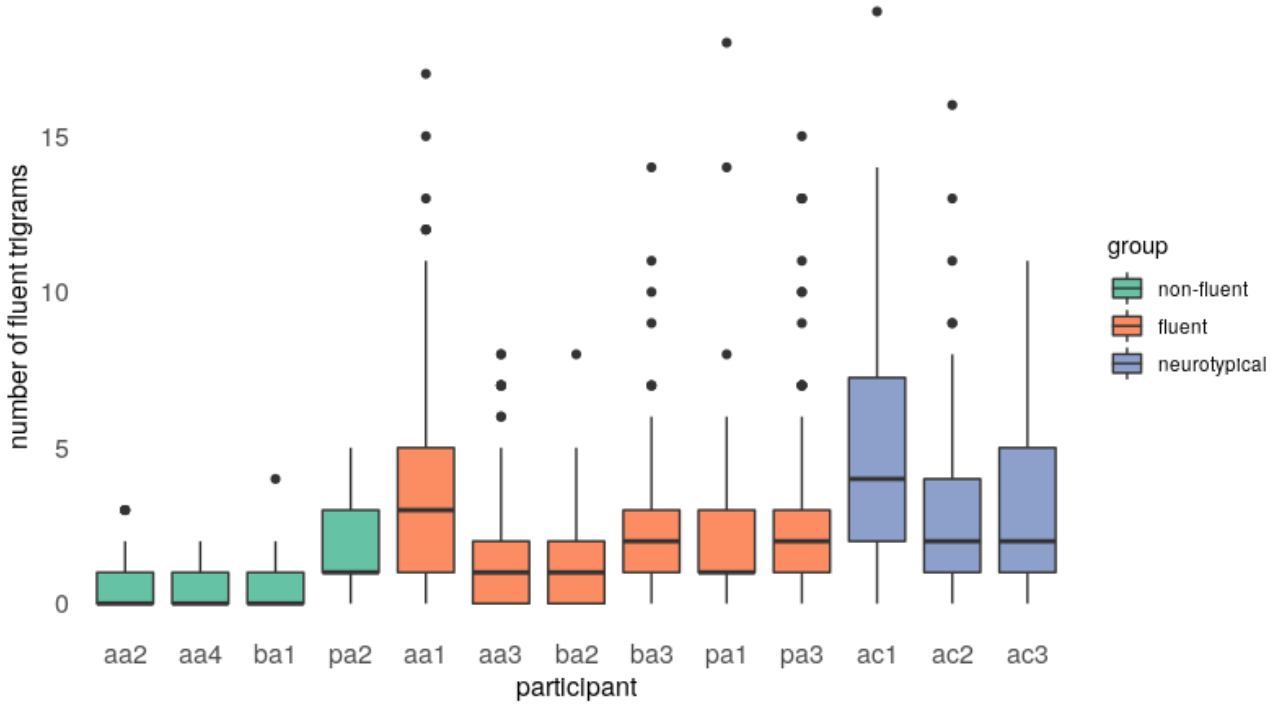

Figure 6: Boxplots showing the distribution of fluent trigrams. Only c-units with a minimum length of three words were included.

4.2.1.3 Disfluent chunks

The ratio of disfluent chunks to fully-formed words produced across c-units is summarized in Figure 7. The non-fluent participants have comparable results here with wider ranges of aa2, aa4, and ba4 compared to ba1 best explained by a difference in MLU. We also see aa3, ba2, ba3, and pa3 emerge as a subgroup among the fluent participants with relatively comparable, average MLUs and a higher frequency of disfluencies. One thing that should be noted is that the disfluent chunks were not differentiated qualitatively with respect to the number and type of disfluency such that a short pause is qualitatively the same chunk as a long pause accompanied with hesitations and word fragments as shown in 9. However, when these differences are taken into account the result of the group comparison remains the same.

9. Quantitatively different disfluency chunks were treated as qualitatively identical when counting disfluencies within c-units:

    a. *a neviděl že tam je 0.2 lev* 'and he didn't see there was 0.2 a lion' (ba2: 17), counted as one disfluency chunk in the c-unit

    b. *a potom dyž panička přišla tak 0.2 h- 0.2 ho nechal* 'and then when the his owner 0.2 l- 0.2 left him alone' (ba2: 65), counted as one disfluency chunk in the c-unit

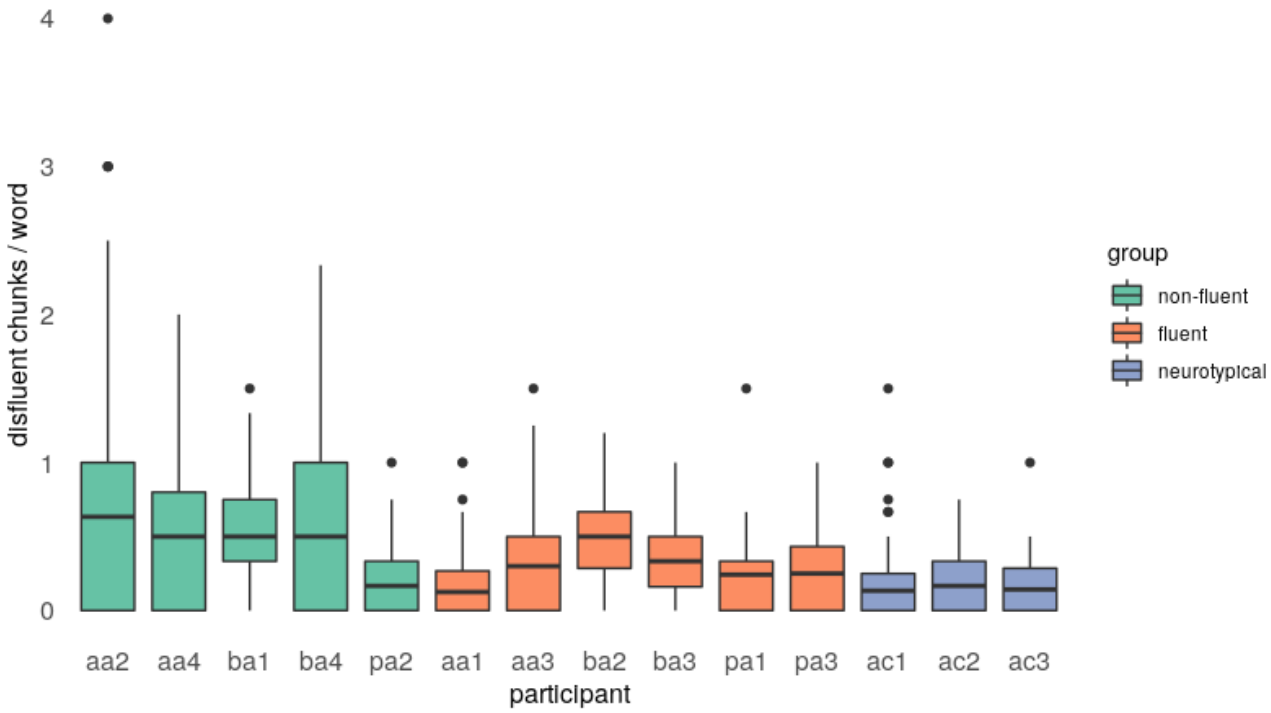

Figure 7: Boxplots showing the distribution of the ratios of disfluent chunks per words.

4.2.2 Productivity

The summary of individual measures of productivity for each participant is provided for reference in Table 12. Particular measures and individual differences are discussed below.

participant	group	zTTR	proportion multiclause	number of modifiers mean (sd)	V / c-unit	N : N+V mean (sd)	log lemma frequency (oral) mean (sd); median (MAD)	proportion high frequency lemma	proportion content words	proportion function words	proportion keywords
aa1	fluent	0.587	0.432	0.932 (1.352)	1.272	0.407 (0.346)	5.371 (3.214); 5.148 (3.52)	0.550	0.474	0.449	0.636
aa3	fluent	-1.495	0.154	0.589 (1.04)	1.030	0.277 (0.373)	6.74 (2.839); 6.571 (3.297)	0.731	0.420	0.432	0.394
ba2	fluent	-0.629	0.295	0.377 (0.753)	1.180	0.362 (0.28)	5.654 (3.334); 5.481 (3.662)	0.596	0.469	0.492	0.697

ba3	fluent	-0.430	0.363	0.763 (1.222)	1.149	0.383 (0.321)	5.755 (3.523); 5.504 (4.17)	0.601	0.459	0.417	0.606
pa1	fluent	-0.087	0.373	0.69 (1.113)	1.115	0.315 (0.302)	5.689 (3.079); 5.223 (3.849)	0.570	0.471	0.396	0.697
pa3	fluent	-1.105	0.426	1.019 (1.421)	1.377	0.23 (0.317)	6.587 (3.248); 6.577 (3.441)	0.738	0.407	0.440	0.485
aa2	nfluent	1.013	0.014	0.22 (0.643)	0.616	0.543 (0.438)	5.29 (3.45); 4.493 (3.684)	0.494	0.521	0.304	0.455
aa4	nfluent	1.381	0.051	0.377 (0.762)	0.468	0.61 (0.406)	5.583 (2.684); 5.193 (2.567)	0.640	0.621	0.232	0.333
ba1	nfluent	1.106	0.194	0.561 (0.923)	0.805	0.601 (0.284)	4.514 (2.66); 4.563 (2.389)	0.490	0.623	0.321	0.576
ba4	nfluent	-2.090	0.000	0.09 (0.388)	0.135	0.84 (0.342)	3.295 (2.102); 3.319 (2.652)	0.286	0.467	0.046	0.333
pa2	nfluent	-0.667	0.157	0.49 (0.784)	0.941	0.422 (0.34)	5.906 (3.155); 5.443 (3.81)	0.590	0.534	0.407	0.364
ac1	nbd	0.096	0.611	1.343 (1.799)	1.472	0.425 (0.274)	5.215 (3.058); 5.178 (3.417)	0.577	0.480	0.439	0.697
ac2	nbd	-1.537	0.442	0.704 (1.292)	1.213	0.379 (0.252)	5.292 (3.221); 4.911 (3.672)	0.552	0.535	0.380	0.848
ac3	nbd	-0.222	0.369	0.571 (0.911)	1.143	0.457 (0.26	4.415 (2.954); 4.143 (2.834)	0.424	0.528	0.417	0.788

	1)

Table 12: Summary of the measures of productivity.

Table 13 summarizes the productivity measures with group and post-hoc pairwise comparisons. We see that only one measure failed to detect any group differences, viz. the proportion of content lemmas with a cumulative frequency of more than 100 pmw in oral. Mean log-transformed lemma frequency of content words was significant in the Kruskal-Wallis test but the post-hoc comparisons were not significant. The mean number of modifiers, mean number of verbs per c-unit, mean ratio of nouns to nouns and verbs in c-units and the proportions of content and function words were found to differentiate between the fluent and non-fluent participants with aphasia, while a difference in zTTR as well as the proportion of multiclause units and story-relevant vocabulary was found between non-fluent speakers with aphasia and neurotypical participants. The individual measures and differences between participants are discussed below.

variable	Kruskal-Wallis (n = 14, df = 2)	pairwise comparison
zTTR	$\chi^2 = 6.94$; $p = 0.0312$; $\eta^2 = 0.449$	nbd x non-fluent (z = -2.55; p = 0.032
mean modifiers	$\chi^2 = 7.55$; $p = 0.0229$; $\eta^2 = 0.505$	fluent x non-fluent (z = -2.408; p = 0.0481); nbd x non-fluent (-2.269; p = 0.0481)
proportion multiclause units	$\chi^2 = 8.741$; $p = 0.0126$; $\eta^2 = 0.613$	nbd x non-fluent (z = -2.815; p = 0.0146)
mean V / unit	$\chi^2 = 9.257$; $p = 0.0098$; $\eta^2 = 0.66$	fluent x non-fluent (z = -2.566; p = 0.0265); nbd x non-fluent (-2.619; p = 0.0265)
mean N to N+V	$\chi^2 = 9.455$; $p = 0.0089$; $\eta^2 = 0.678$	fluent x non-fluent (z = 3.066; p = 0.0065)
proportion high frequency tokens	$\chi^2 = 3.773$; $p = 0.152$	
mean log frequency	$\chi^2 = 6.364$; $p = 0.0415$; $\eta^2 = 0.397$	n.s.
proportion content words	$\chi^2 = 7.354$; $p = 0.0253$; $\eta^2 = 0.487$	fluent x non-fluent (z = 2.448; p = 0.0431)
proportion function words	$\chi^2 = 8.284$; $p = 0.0159$; $\eta^2 = 0.571$	fluent x non-fluent (z = -2.867; p = 0.0124)
proportion keywords	$\chi^2 = 9.175$; $p = 0.0102$; $\eta^2 = 0.652$	nbd x non-fluent (z = -2.984; p = 0.0085)

Table 13: Group comparisons of the measures of productivity.

4.2.2.1 Lexical diversity

This corpus-based lexical diversity measure was the only one where the three different tasks contained in the subcorpus were taken into account. The rationale behind this decision was that this measure might be considerably more sensitive precisely to the fact that the subcorpus is assembled from three different discourses within the interviews. Task based measures were averaged over the three tasks and these values were used for comparison. The average scores are plotted in Figure 8 and measures for individual tasks are shown in Figure 9.

Note that values around zTTR = 0 are similar to corpus samples of corresponding length, while negative values indicate lower and positive values greater lexical diversity. We see a clear difference

between the neurotypical speakers and the participants with aphasia in Figure 8. Only aa1 has a positive value and patterns with the neurotypical group. ba1, pa1, and pa2 have values around zero, i.e. comparable to spoken Czech as represented by ORAL. We also see a group of aa2, aa3, ba3, and pa3 with markedly lower lexical diversity. These are caused by two different language profiles. The production of the nonfluent participants aa2 and ba4 is characterized by low overall numbers of tokens and frequent repetitions following word finding difficulties. Participants aa3 and pa3 on the other hand produced a large number of tokens but their speech is characterized by a frequent use of pronouns and discourse markers, i.e a lower information load. These two very different profiles result in similar zTTR scores, given the nature of the measure. Figure 9 explains the comparatively "less low" lexical diversity in participant aa4 who performed "around-average" on the CHAPLIN and LADA task, but scored much lower on the COMIC task. A similar pattern can be seen in ba1 who performed similarly to the neurotypical speakers on CHAPLIN and LADA, but has a low zTTR value for COMIC. The performance of the other participants is relatively stable across tasks.

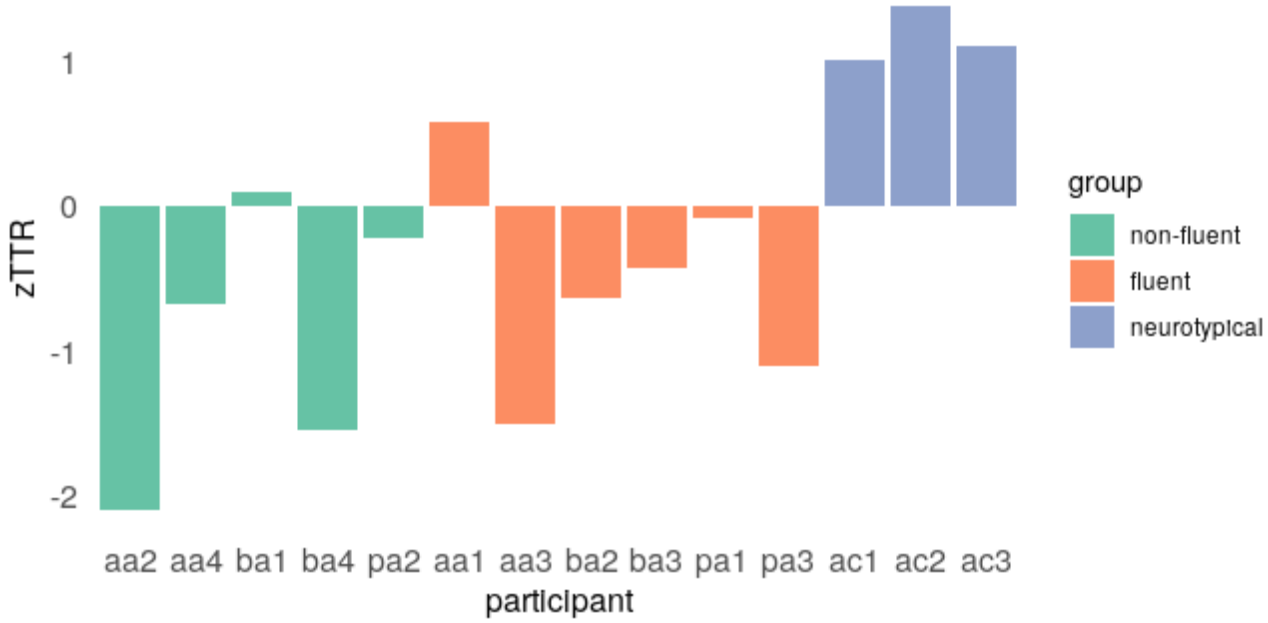

Figure 8: Average zTTR scores across tasks.

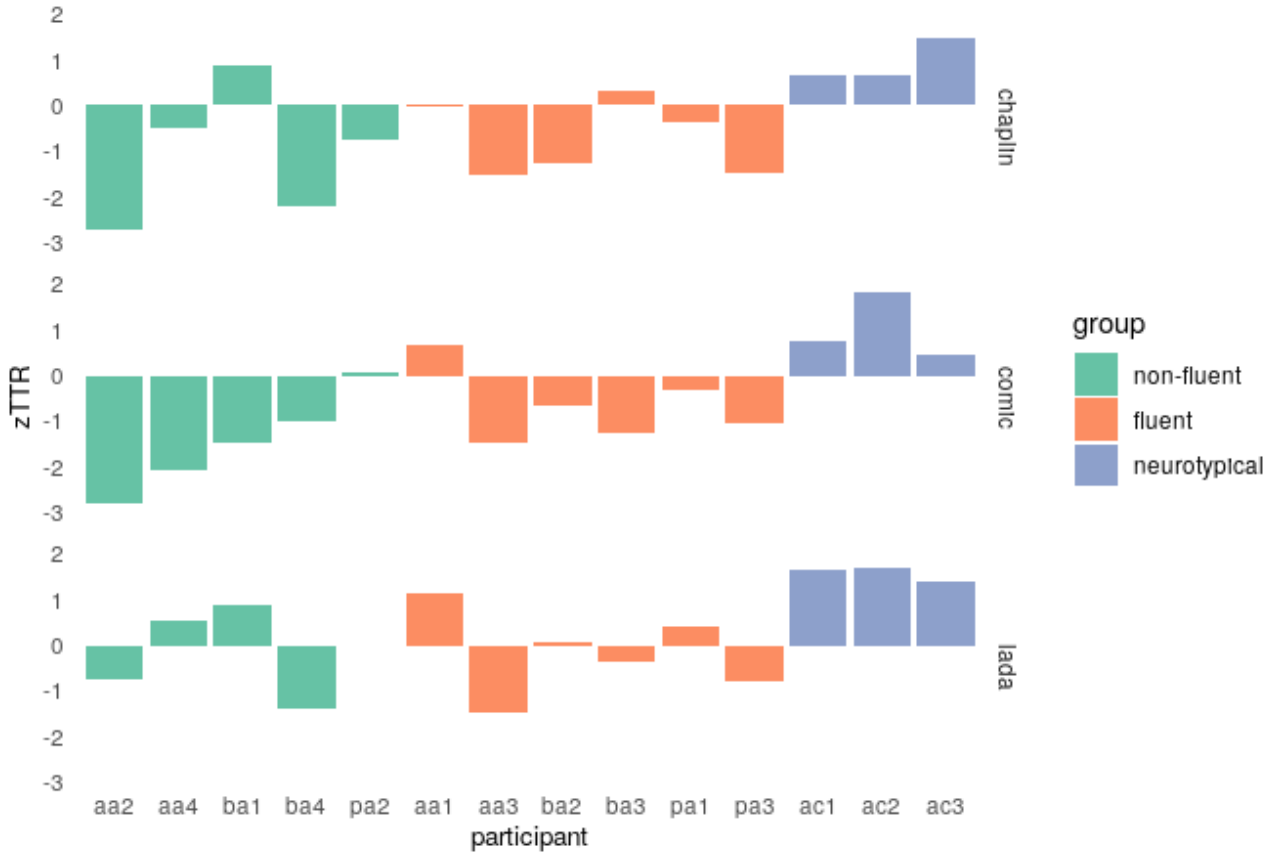

Figure 9: zTTR values for individual tasks.

4.2.2.2 Number of modifying elements

The overall number of modifying and backgrounded elements was relatively low for all participants. This can be seen from the fact that the third quartile value was equal to 1 for almost all the participants with the exception of aa2 and aa4 (q3 = 0) on the one hand and pa3 and ac1 (q3 = 2) on the other. While the pairwise comparison suggests a difference both between fluent and non-fluent as well as neurotypical and non-fluent participants, Figure 10 showing individual mean values suggests that these differences are driven only by the performance of participants aa2 and ba4. When these two participants are not considered, there is no clear group pattern. We only see that aa1 and pa3 pattern with ac1.

A conclusion could be made that these differences may be driven by individual narrative styles rather than an adaptive strategy that might be expected in persons with aphasia who could be predicted to try to avoid communicating any non-essential information that could increase processing load. There is also a possibility that the tasks, in particular CHAPLIN and COMIC that are more focused on conveying a story, are simply not appropriate for the elicitation of "added information". The use of visual stimuli can further contribute to that, as participants may not feel the need to communicate information about, say, the colors of objects, as this information is visually available to both interlocutors. Figure 11 shows the average number of modifiers used by participants in the

three tasks. We see that some of the participants did use more modifiers in the LADA task, as might be expected, given that they were asked to describe the picture in as much detail as possible. However, the general pattern, or rather, no pattern remains the same. Overall, this measure seems not to be very informative, at least in the context of the present sample.

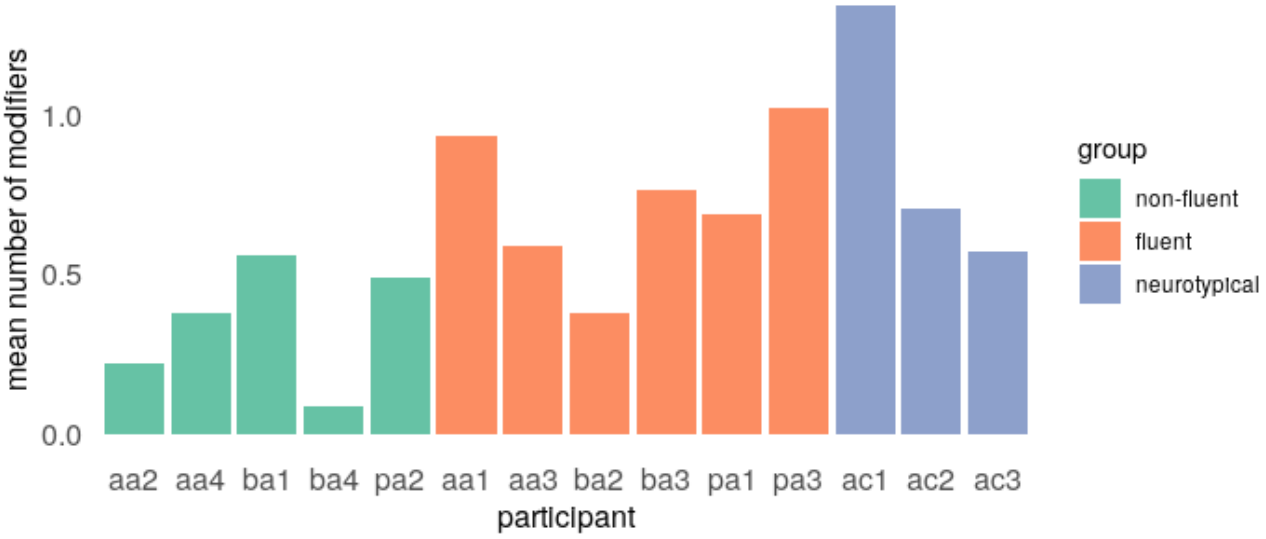

Figure 10: Mean number of modifying elements produced across c-units.

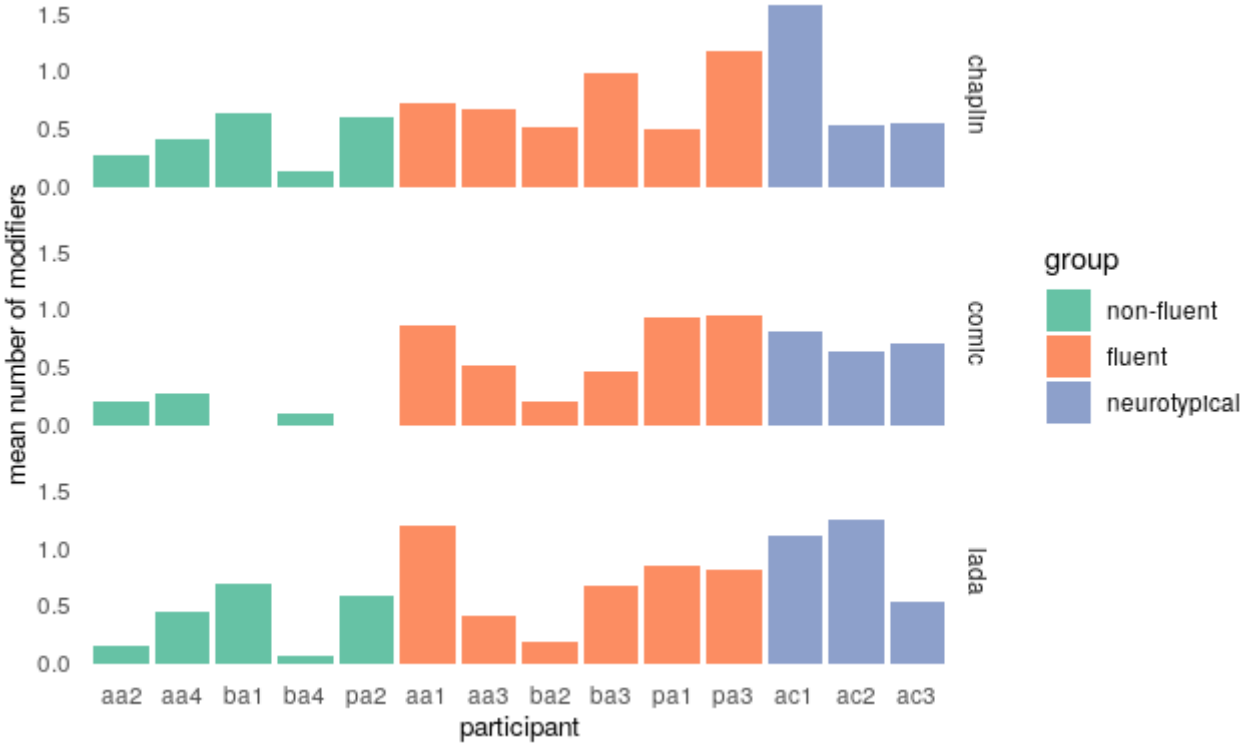

Figure 11: Mean number of modifying elements produced in individual tasks.

4.2.2.3 Multiclause units

The overall occurrence of multiclause c-units was not very high, which is to be expected in spoken language which is generally characterized by less subordination compared to the written domain. However, we do see a clear pattern where the speakers with non-fluent aphasia produce almost no complex clauses with subordination, as evidenced by Figure 12. Both aa2 and aa4 produced only a single c-unit with subordination shown in 10. Note that aa2's token contains the chunk *myslel si že* '(he) thought that' which can be expected to be quite frequent in the spoken language, the complement clause is left unfinished. The relative clause produced by aa4 is also abandoned. Note also that there is a good deal of individual differences visible also among the neurotypical participants. While ac1 produced 37 % of multiclause c-units, ac2 and ac3 produced around 20 % of clauses with subordination.

10. Multiclause c-units produced by aa2 and aa4

    a. *a 1 (no) 0.2 a myslel si že 0.5 H 0.5 byl 0.5 akovej* 'and 1 (well) 0.2 and he was thinking that 0.5 <hes> 0.5 he was 0.5 such' (aa2: 104)

    b. *ano vyše? mu 0.2 na po(l)moc 1 hloj- hol- holka: která 0.2 byla 1 H 5 tak 3 hod- hoka holka H* 'yes came to him 0.2 to help 1 griw- gir- girl who 0.2 was 1 H 5 so 3 giw- gil girl H' (aa4: 63)

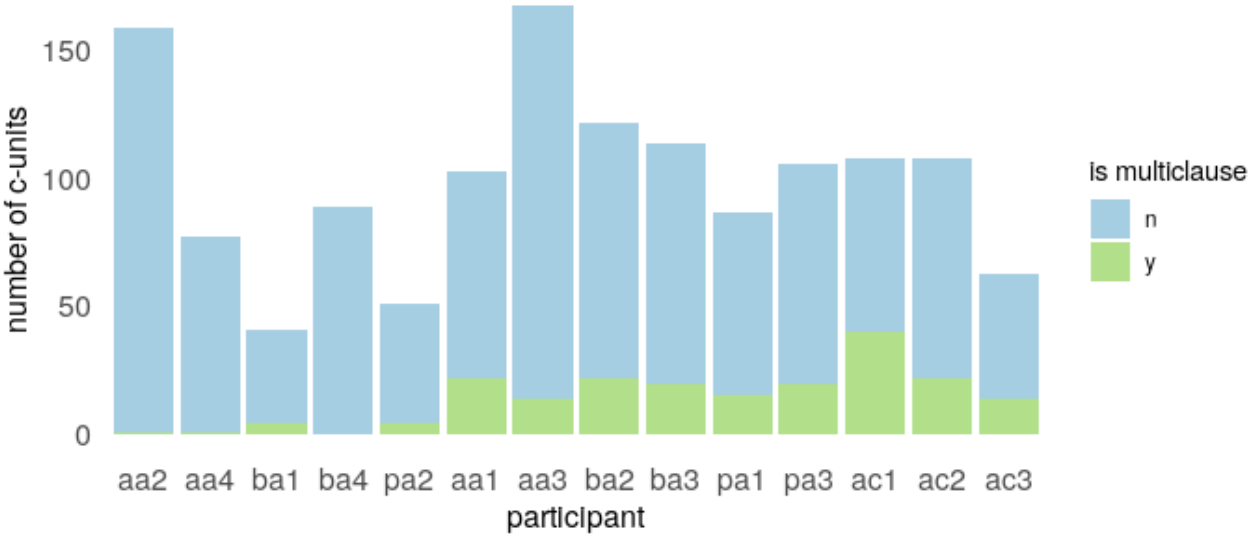

Figure 12: Summary of c-units with a single predication and of multiclause c-units.

4.2.2.4 Verbs and nouns

The two measures used to assess the production of verbs and nouns are summarized in Figure 13. Given the diagnoses of the patients it was expected that the non-fluent participants will on the whole produce less verbs as has been long established in the literature. Figure 13 shows the mean number of verbs in the upper panel and the mean ratio of nouns to the total number of nouns and verbs across c-units. Note that multiclause units are also included in these counts. The upper panel shows that participants are grouped based on whether they produced a verb in the majority of c-

units or not. This clearly differentiates between the non-fluent participants and the fluent and neurotypical group. The only exception here is pa2 whose production is comparable to the fluent group. This pattern is also mirrored in the lower panel where we see a higher proportion of nouns in participants aa2, aa4, ba1, and ba4. aa3 and pa3 are interesting with regard to noun to noun and verb ratio as they show relatively lower proportions of nouns compared to other participants. This is in line with the pattern mentioned above that concerns the production of fully-formed, relatively long sentences with lower content value characterized by fewer nouns. A detailed discussion of verb production is provided in section 4.3.1.

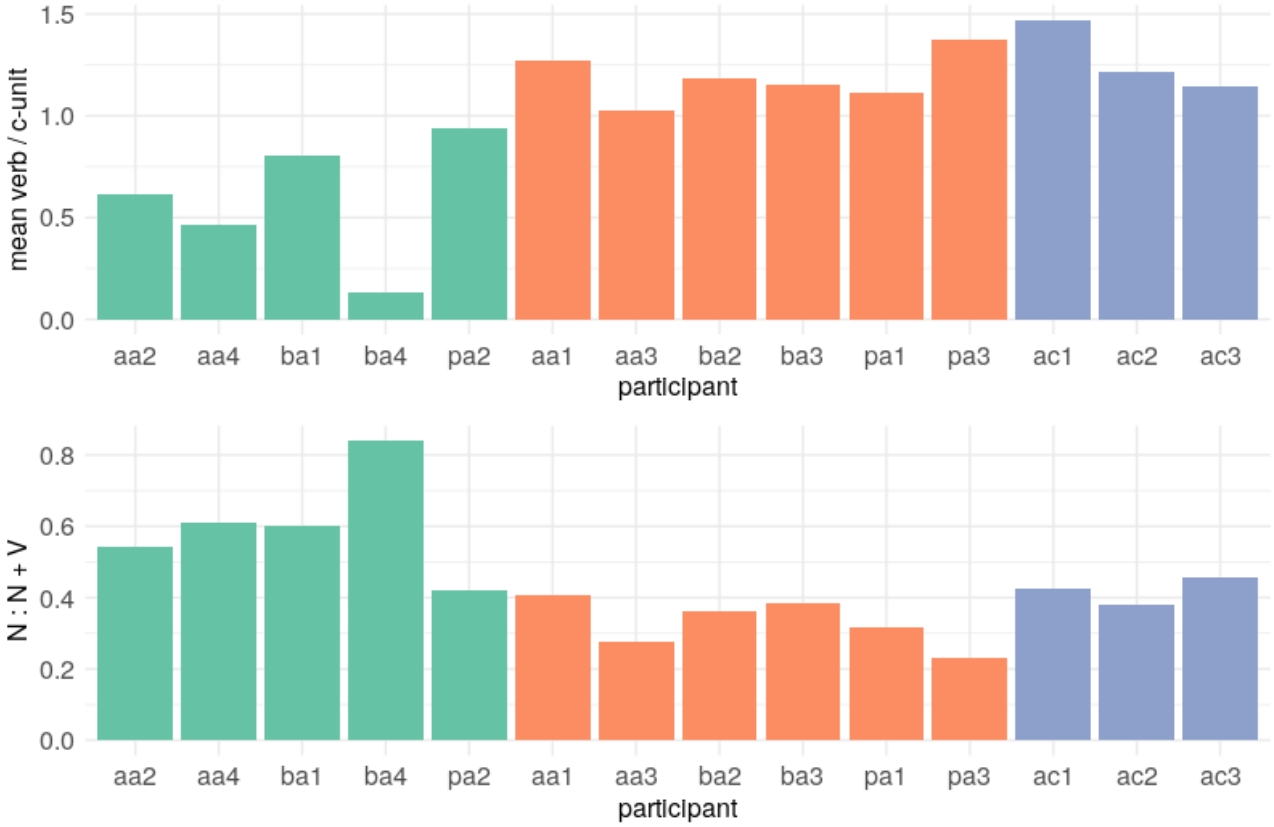

Figure 13: Production of verbs and nouns in the subcorpus. The upper panel shows mean number of lexical verbs per c-unit, the lower panel shows the mean ratio of nouns to the number of nouns and verbs.

4.2.2.5 Frequency

The rationale behind the two frequency measures used in the analysis was that participants with lexical retrieval problems might rely more on high frequency words that are easier to retrieve. The threshold for the high frequency band was 100 pmw. The raincould plot (Allen et al. 2021) in Figure 14 shows the distribution of all content lemma frequencies (log transformed) for individual participants.[41] No group patterns are visible as also suggested by the post hoc comparison. However, two observations deserve a comment.

41 Raincloud plots are a data visualization technique that combines the advantages of boxplots and density plots in a single figure that provides a detailed information about the representation of the data.

First, we see that a comparatively low number of (very) high frequency lemmas was used by the non-fluent participants aa2 and ba4 (median frequency around 90 and 30 pmw respectively). This can be explained by the fact that these participants produced short utterances typically consisting of a noun and, in the case of aa2, a predicate. The production of these nouns and verbs was driven by the stimuli used in the tasks that necessarily do not depict highly frequent referents (e.g. *lvice* 'lioness' or *zabijačka* 'hog killing') and the tokens were in many cases produced in collaboration with the administration. Conversely, we see that participants aa3 and pa3 used more high frequency lemmas (median frequency of 735 pmw for both participants). This corresponds to their general language profile discussed throughout this section characterized by less specific and less informative language.

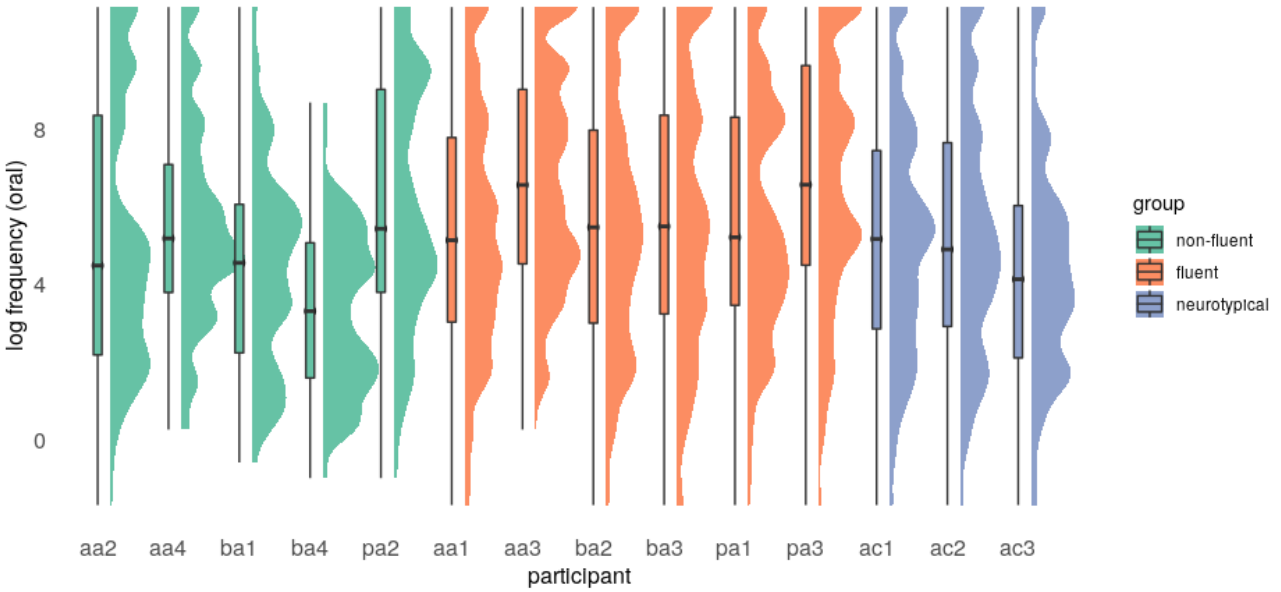

Figure 14: Raincloud plot with the distribution of log-transformed oral frequencies of content word tokens. Boxplots are combined with density plots.

4.2.2.6 Content and function words

The proportion of content and function words has been a standard measure in aphasiology and is motivated by the observation that non-fluent speakers, in particular persons with Broca's aphasia, encounter difficulties in the production of closed-class words. Czech is typologically different from the languages frequently studied in aphasia research in that a substantial proportion of grammatical morphemes are affixes and their omission would oftentimes result in nonwords, which makes paragrammatic errors more frequent (e.g. Lehečková 2001). Nonetheless, the expected pattern is visible in Figure 15. In particular, the non-fluent participants aa2, aa4, and ba4 have a considerably lower proportion of function words; a relatively lower proportion is also visible in ba1. On the other hand, participant pa2 who is also classified clinically as non-fluent patterns with the other speakers. Note also in the upper panel that speakers aa3 and pa3 have a slightly lower proportion of content words which is consistent with their general language profile, as their production included a high number of discourse markers, pronouns, and conjunctions.

Figure 15: Proportion of content words (upper panel) and function words (lower panel).

4.2.2.7 Story relevant vocabulary

The use of task specific vocabulary may be taken as an indicator of discourse informativeness and the degree of lexical retrieval problems. Participants with a production pattern characterized by more severe word finding difficulties are expected to produce a lower number of core vocabulary lemmas. Similarly, the emerging language profile of participants aa3 and pa3 characterized by reliance on high frequency content words and pronouns as well as on overuse of discourse markers can be expected to produce a lower proportion of story relevant vocabulary.

Story relevant lemmas obtained for the three tasks in the questionnaire (cf. section 3.5) are given in Table 14. Frequency describes the number of participants in the corpus who produced the given lemma.

chaplin		comic		lada	
lemma	frequency	lemma	frequency	lemma	frequency
lev ‘lion’	14	kočka ‘cat’	12	dítě ‘child’	9
klec ‘cage’	8	trouba ‘stove’	10	zabijačka ‘hog killing’	11
dveře ‘door’	8	váza ‘vase’	10	obrázek ‘picture’	5
Chaplin	11	myš ‘mouse’	12	prase ‘pig’	11
probudit (se) ‘wake’	6	stůl ‘table’	4	hrnec ‘pot’	9
otevřít ‘open’	8	jídlo ‘food’	1	pes ‘dog’	8
tygr ‘tiger’	10	květina ‘flower’	4	řezník ‘butcher’	9
omdlít ‘faint’	9	žena ‘woman’	4	Josef	3
pes ‘dog’	11	rozbít (se) ‘break’	8	nést ‘carry’	8
Charlie	2	skočit ‘jump’	4	voda ‘water’	11
dostat se ‘escape’	5				
snažit se ‘try’	5				
spát ‘sleep’	11				

Table 14: The use of story-relevant lemmas across tasks; frequency shows the number of participants who produced the lemma.

As may be seen in Figure 16 which breaks down the proportion of story relevant lemma uses by task, participants on the whole produced a substantial number of the words that were obtained in the questionnaire. When we look at the non-fluent group, aa4 produced less than 50 % of the expected vocabulary in all three contexts. ba4 and pa2 had a low number of keywords in two of the three tasks. As for the other two non-fluent participants, aa2 was relatively unsuccessful on CHAPLIN and ba1 has a production comparable to the fluent speakers. This can be interpreted such that aa4, ba4, and pa2 have more impaired lexical retrieval. We also see the expected pattern for participant aa3 was less successful in core vocabulary production. On the other hand, pa3 produced a lower proportion of story relevant lemmas only in COMIC.

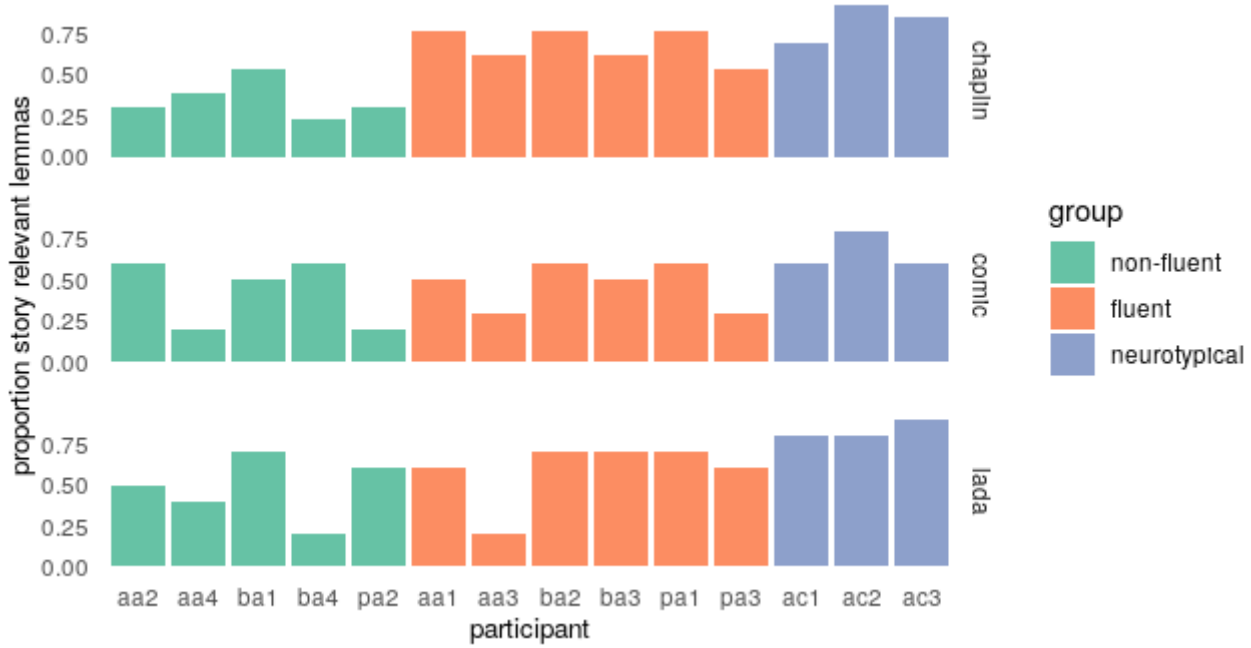

Figure 16: Proportions of story relevant lemmas produced in individual tasks.

4.2.3 Well-formedness

The summary of individual measures of well-formedness for each participant is provided for reference in Table 15. Particular measures and individual differences are discussed below.

participant	group	proportion fragments	proportion interventions	proportion retracing	proportion c-units with errors
aa1	fluent	0.049	0.093	0.136	0.078
aa3	fluent	0.256	0.398	0.268	0.429
ba2	fluent	0.074	0.056	0.131	0.148
ba3	fluent	0.175	0.097	0.368	0.158
pa1	fluent	0.034	0.148	0.161	0.149
pa3	fluent	0.189	0.397	0.283	0.302
aa2	nfluent	0.390	0.500	0.289	0.258
aa4	nfluent	0.273	0.402	0.169	0.429
ba1	nfluent	0.122	0.146	0.024	0.171
ba4	nfluent	0.584	0.298	0.000	0.573

pa2	nfluent	0.098	0.547	0.078	0.157
ac1	nbd	0.056	0.086	0.306	0.037
ac2	nbd	0.009	0.022	0.102	0.065
ac3	nbd	0.000	0.134	0.127	0.000

Table 15: Summary of the measures of well-formedness.

Table 16 summarizes the well-formedness measures with group and post-hoc pairwise comparisons. Three of the four measures used to assess well-formedness of c-units suggested a group difference and following pairwise comparisons revealed significant differences between participants with non-fluent aphasia and neurotypical speakers. The proportion of c-units with repetitions and reformulations (retracing) did not detect any group differences.

variable	Kruskal-Wallis (n = 14, df = 2)	pairwise comparison
proportion sentence fragment	$\chi^2 = 7.154$; $p = 0.028$; $\eta^2 = 0.469$	nbd x non-fluent (z = 2.662; p = 0.0234)
proportion interventions	$\chi^2 = 6.714$; $p = 0.0348$; $\eta^2 = 0.429$	nbd x non-fluent (z = 2.51; p = 0.0363)
proportion retracing	$\chi^2 = 2.417$; $p = 0.299$	
proportion errors	$\chi^2 = 7.796$; $p = 0.0203$; $\eta^2 = 0.527$	nbd x non-fluent (z = 2.785; p = 0.016)

Table 16: Group comparison of the measures of well-formedness.

4.2.3.1 Proportion of sentence fragments

Sentence fragments include c-units abandoned due to unsuccessful retrieval, single word c-units with no predication, and tokens that also occurred due to retrieval problems that resulted in clauses that do have a predication but the resulting structure is either incomplete, e.g due to a missing direct object of a transitive predicate, or the speaker attempts a reformulation that is unsuccessful, rendering the structure ungrammatical.

Figure 17 shows the proportion of sentence fragments in relation to the total number of c-units produced by participants. We see that the majority of ba4's production were fragments. These were mostly single word utterances or listings of several nouns. aa2 and aa4 also produced a high number of fragments that were mostly single word utterances or structural fragments of the type illustrated in 11.a and b. The other participants in the non-fluent group produced less fragments but the propor-

tion is still higher than for some of the more fluent speakers. We see again that aa3 and pa3 perform in a way comparable to the non-fluent group in that their frequent retrieval problems resulted in a high number of abandoned c-units and structurally ill-formed c-units. A similar pattern is also seen in ba3.

11. Single word utterance and sentence fragment

    a. *tady 1 závora* 'there 1 a latch' (aa2: 101), this is an example of a c-unit without a full predication; the missing verb is most likely *be*

    b. *a je 0.2 za pět minut 0.5* 'and it's 0.2 five minutes to' (ba2: 142), the speaker was trying to tell the time shown in one of the stimulus pictures, the utterance was left unfinished for no clear reason

Figure 17: Proportion of sentence fragments (upper panel) and number of fully formed and fragment c-units (lower panel).

4.2.3.2 Interventions

The number of prompts and interventions produced during the interviews by the administrator to collaborate with participants on repairs and during word finding difficulties was overall relatively high for the non-fluent group with the exception of ba1, as can be seen in Figure 18. Note that pa2 who in many of the discussed measures differed from the other non-fluent speakers, has a high proportion of administrator interventions. This was driven by the fact that pa2 had a very hesitant production pattern with long silent pauses indicative of a slow-down in processing. These were taken

by the administrator as a sign of retrieval problems. This was also commented upon by the participant during the interview:

12. *H 0.5 dyž von počká 0.2 tak mu to řeknu / ale jak von H b- bude čeka- bude p- chvátat 0.2 tak sem v prdeli* 'H 0.5 if he waits 0.2 I am able to say it / but if he H w- will wai- will w- rush 0.2 then I'm screwed' (pa2: 182-183)

We again see a high proportion of interventions in aa3 and pa3. This pattern is related to the number of sentence fragments discussed above and was caused by collaborative repairs and prompts during word finding difficulties. This is particularly the case for participant aa3. In a representative situation the participant was prompted with the target word when lexical retrieval failed and his relatively highly impaired repetition caused by his conduction aphasia oftentimes resulted in a series of repeated prompts during which the participant was trying to repeat the target.

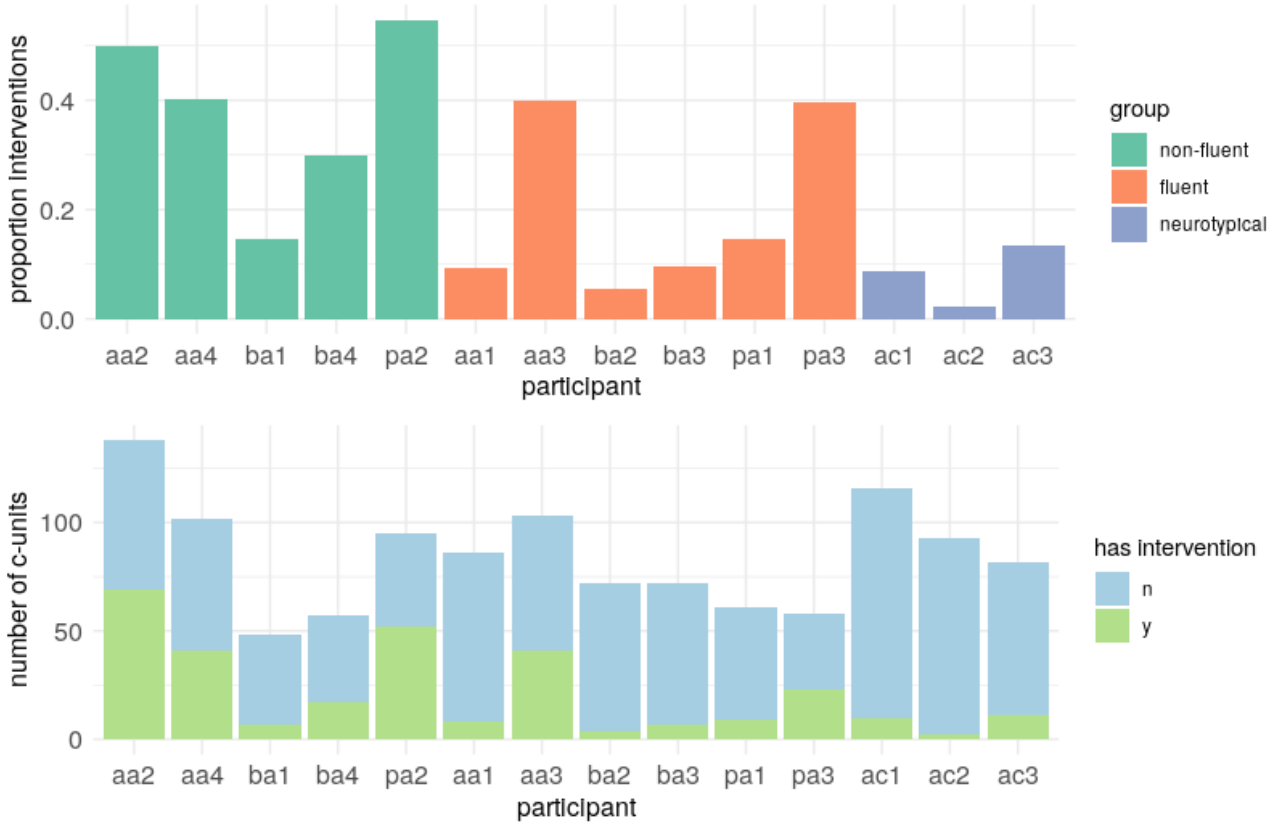

Figure 18: Proportion of interlocutor interventions in relation to the number of c-units (upper panel); number of c-units according to whether they were followed by interventions (lower panel).

4.2.3.3 Retracing

While the comparison of the proportion of c-units containing repetitions and reformulations did not reveal any group differences, there are interesting individual patterns. Figure 19 shows that participants aa3, ba3, and pa3 had a high proportion of retracing in their production which again reflects the same problems discussed with regards to sentence fragments and administrator interventions for aa3 and pa3. ba3 has a slightly different general language profile mainly driven by frequent repetitions which are related to a symptom that was particular to this participants, i.e. frequent stuttering

that was also described in the assessment of the participant's SLT. 13 is representative of ba3's production. We see several word fragments and repetitions. aa2 is the only participant from the non-fluent group with a high number of retracing units. This reflects a specific retrieval strategy of the participant. In a representative situation, he produced a word that was close to the target word and repeated it while trying to reformulate it and retrieve the target expression. In context where the continuation of an utterance caused word finding difficulties, I interpret this as a strategy to keep the train of thought and help the participant in finding the desired continuation of the utterance. Participant ac1 is the reason why no group differences could be detected as we see that he produced a considerable proportion of reformulations. The source of this was in the CHAPLIN task where the participant was trying to remember a part of the story which caused a distortion in his production.

13. *t- 0.5 tomu te vy- té musi tady H 0.2 udělat vo- 0.2 volno prože n- nese vi- pivo* 'h- 0.5 him her ma- to her they must H 0.2 make w- 0.2 way here becuase she's b- bringing wi- beer'(ba3: 243)

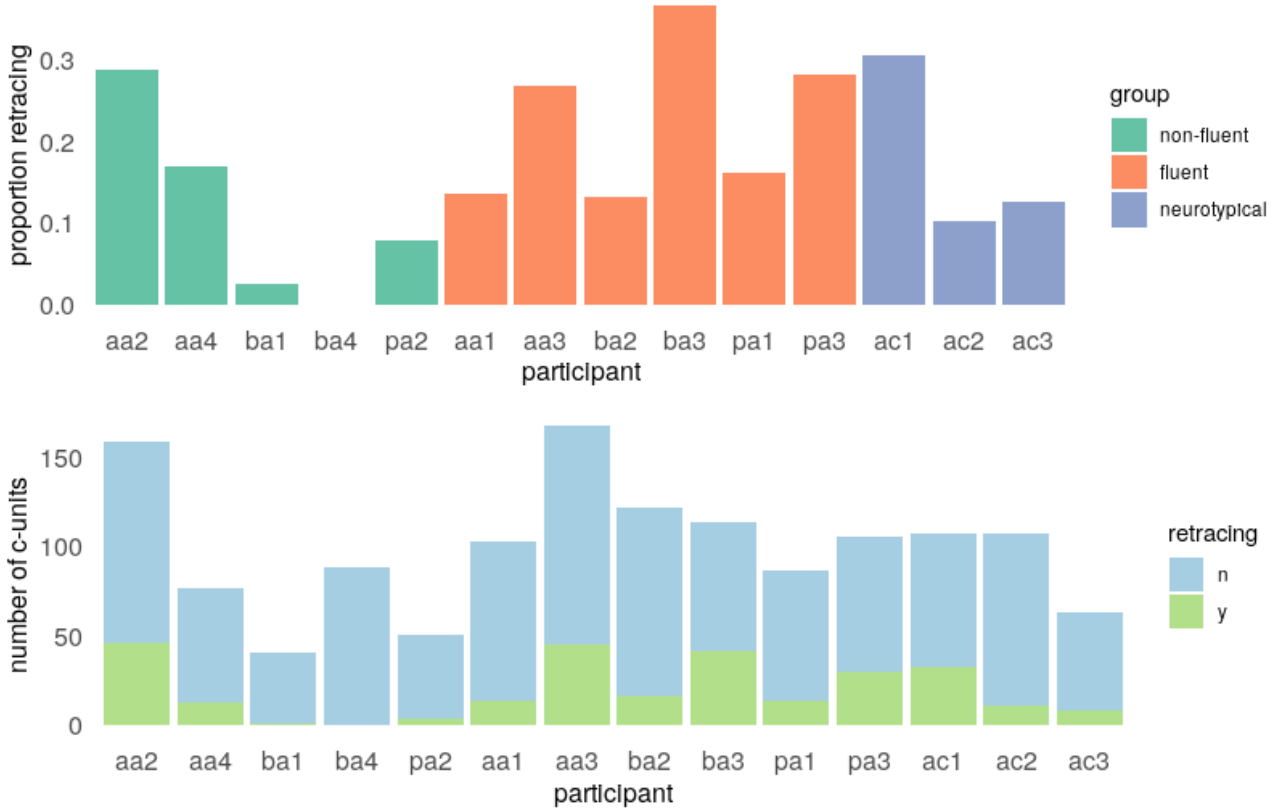

Figure 19: Proportion of c-units with retracing (upper panel) and number of c-units with and without retracing (lower panel).

4.2.3.4 Proportion of errors

The proportion of c-units produced with an error of any kind can be taken to indicate differences in the levels of severity within the fluent and non-fluent group. Participants aa2, aa4, and ba4 produced a higher number of errors than ba1 and pa1, as shown in Figure 20. Figure 21 breaks down the errors into four categories of phonemic paraphasias, word finding difficulties, paragrammatic and agrammatic errors, and discourse level errors.

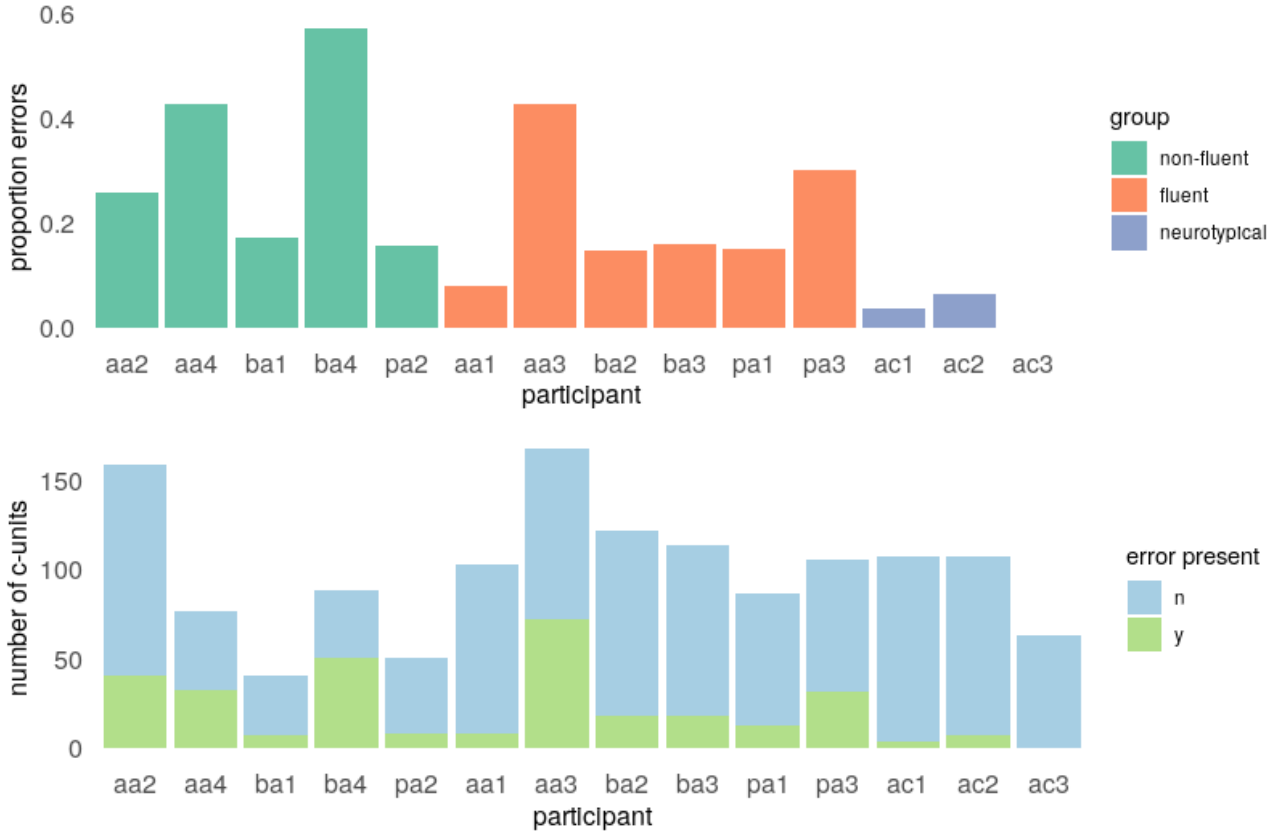

Figure 20: Proportion of c-units with errors (upper panel); number of c-units produced with and without errors (lower panel).

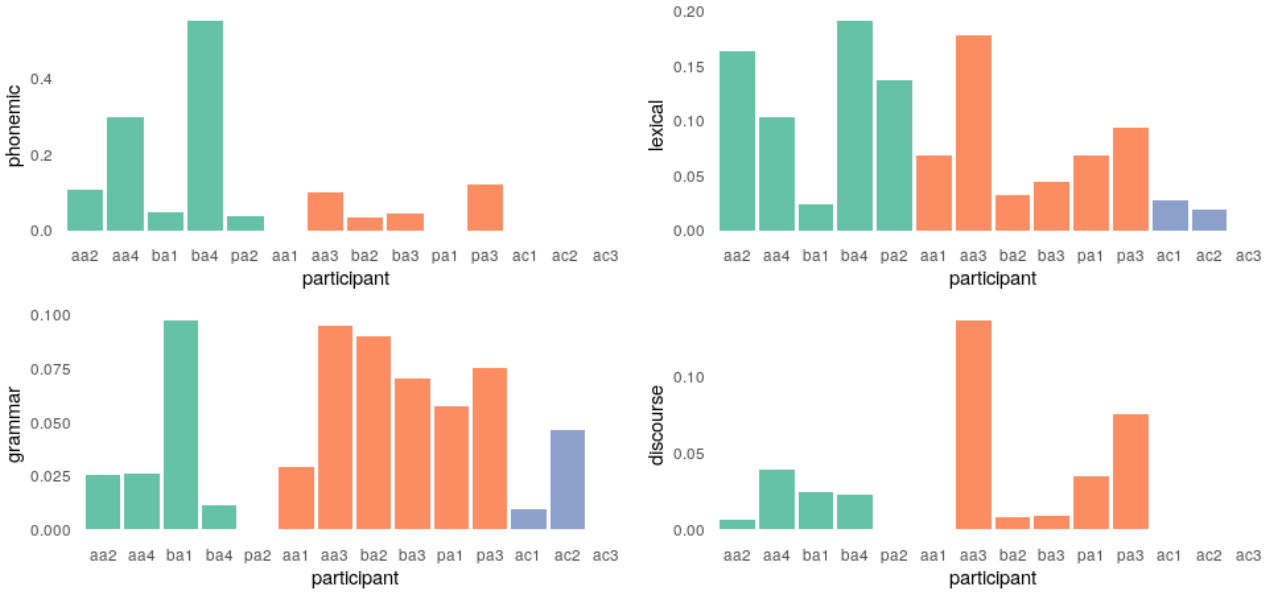

Figure 21: Proportion of c-units produced with phonemic errors (upper left hand panel), lexical retrieval errors (upper right hand panel), grammatical errors (lower left hand panel) and discourse level errors (lower right hand panel).

We see a high proportion of phonemic paraphasias and word finding errors in aa2, aa4, and ba4. It should be noted that aa2, aa4, and ba4 have oral apraxia which explains the high number of phonemic errors. aa3 and pa3 are another group with a high proportion of errors. These are mostly lexical retrieval errors. However, we also find a relatively high number of discourse level errors compared to other participants. These are mostly c-units with coherence problems caused by pronouns with no or unclear antecedents.

Grammatical errors were relatively rare in the sample. A higher number of grammatical errors was produced by aa3, ba1, ba2, ba3, and pa3. This agrammatic and paragrammatic production falls for the most part into two types. First, a number of cases involves a mismatch in gender and/or number. These were frequently repaired to target values as shown in 14.a. A larger number of unrepaired mismatches was only produced by aa3, as illustrated in 14.b. The very few instances of case mismatch produced are in pa1 and pa3's data. These two participants used the accusative word form *ho* 'him-ACC' in the context where the dative *mu* would be expected, as shown in 14.c-d. The intended reading here is 'She will open [the door] for him'. While the accusative used here marks the direct object and theme with the verb *otevřít*, the dative must be used to express the beneficiary. It might be the case that this is an idiosyncratic or dialectal feature. However, given the fact that pa1 produced the expected form in the immediately preceding c-unit and that pa3 is from a different region of Czechia, I would argue that these are paragrammatic errors. 14.d shows one additional case mismatch produced by pa3 in relative pronoun. The nominative form *kerá* 'which' is originally produced which would introduce a subject relative, this is however immediately repaired to the accusative form as the head noun *myšku* 'mouse' is intended as the patient and direct object in the relative clause.

14. Examples of representative grammatical errors in the corpus:

a. *a H snažil se i do vedlejšího 0.2 do vedlejší klece utect H tou H těma dveřma* ‘and H he was also trying to escape to the adjacent-M.GEN 0.2 to the adjacent-F.GEN cage H through the-F.SG.INS H the-F.PL.INS door’ (aa1: 42), note that two repairs were made here, first, the mismatched masculine adjective form is corrected to the appropriate feminine form and then the singular demonstrative is repaired the plural form (door is a plurale tantum in Czech)

b. *a zase 0.2 spala dál 1* ‘and then 0.2 she was sleeping on’ (aa3: 186); the referent here was clearly a lion, a masculine noun in Czech

c. *no 0.5 pak přiběhla 0.5 přiběhlo děvče 0.5 aby mu votevřela 0.5 / vona tam 0.5 <tut> misto aby ho vote- votevřela tak tam padla* ‘and 0.5 then came-F 0.5 came-N a girl 0.5 to open (the door) for him-DAT 0.5 / there she 0.5 <tut> instead of ope- opening him-ACC she dropped down there’ (pa1: 27-28); note also the feminine form corrected to neuter (*děvče* ‘girl’ is grammatically neuter in Czech)

d. *a vona na to přišla že jako je to votevřený a 0.2 že ho votevře* ‘and the realized that like it’s open and 0.2 that she’ll open him-ACC’ (pa3: 55)

e. jo *0.2 aha 0.2 a tadydleta uvidí 0.2 H 0.2 tadydle 0.2 tu myšku kerá k- kerou <sigh> by si chtěla jako sníst 0.2 že by si ho sedla* ‘right 0.2 yes 0.2 and this one sees 0.2 H 0.2 this 0.2 there mouse that-NOM th- that-ACC <sigh> she’d want to eat 0.2 that would at him’ (pa3: 118); note also that the pronoun ho ‘him-ACC’ is ungrammatical here as it refers to the mouse (feminine in Czech), the form *sedla* is paraphasic with the target most likely being *snědla* ‘eat up’

The other group represented by several examples in the data are reflexive pronouns. Several verbs are used in the corpus with the accusative reflexive pronoun form *se* where the dative form *si* is expected, such as 15.a. There are also a few cases where a reflexive pronoun is used with a verb that normally does not appear in the reflexive construction. These occurrences appear in retrieval problem contexts and may be interpreted as blends of the target verb and the retrieved one, such as 15.b where the target might be expected to have been *se zvedl* ‘(he) stood up’, but the verb *vstal* ‘got up’ was produced instead.

15. Substitution errors in reflexive pronouns

a. *za chvíli se 0.2 H 5 (v)šiml že tam je lev* ‘in a while REFL 0.2 H 5 he noticed that there was a lion’ (ba1: 21)

b. *a (ště) ten l- lev se 1 vstal* ‘and (also) the l- lion REFL 1 stood up’ (ba3: 66)

Lastly, there is a small number of cases that were coded as grammatical errors that concern word order. A representative example is shown in 16 and concerns the placement of the accusative pronominal clitic *ho* ‘him-ACC’ which would be expected in the Wackernagel position, i.e. after the first constituent of the clause which is the verb in this case (Wackernagel et al. 2020). While there are a few similar tokens in the subcorpus, it may well be the case that this word order pattern could be per-

fectly acceptable, at least for some speakers.[42] In any case it may be said that it is at least dispreferred.

16. *a 0.5 H 0.2 H 0.2 ho pustila ven 0.5* 'and 0.5 H 0.2 H she him let out 0.5' (ba2: 99)

4.2.4 Conclusion

I conclude this section with an overview of the general language profiles that emerge from the described measures and a discussion of the groups and individual differences and similarities. In order to better understand the overall language profiles of the individual participants and their similarities and differences, I submitted the measures for which significant group differences were found to a hierarchical cluster analysis. The only exception was the mean number of modifiers which upon closer inspection proved to be of little descriptive value for this particular sample.[43]

The measures were scaled and a standard Euclidean distance matrix was computed and a hierarchical clustering algorithm using the Ward's method was performed. A solution with two clusters was selected as best representing the structure of the data using average silhouette width as criterion. The resulting cluster dendrogram is shown in Figure 22. The figure also shows distances between individual participants in two dimensional space.

42 The following is an example of the same pattern from oral: *sem mysim řikala malemu ať to .. hodinu .. se neukazuje a ho nejde ani vidět ani slyšet*. 'I think I told the little one to .. for an hour .. not show himself and make himself not seen or heard', where the expected order would be ... *a nejde ho...* with the accusative pronominal clitic placed after the verb. The fact that we find a similar construction is not surprising, since aphasic errors are not qualitatively different from what we find in neurotypical spontaneous speech as discussed, for instance, by Lehečková (2008).

43 To reiterate, the measures used for the clustering were the following: MLU, mean number of fluent trigrams, number of disfluent chunks per word, average zTTR scores from the three task, the proportion of multiclause c-units, mean number of verbs per c-unit, mean ratio of nouns to nouns and verb, the proportion of produced task story relevant words, the proportion of sentence fragments, the proportion of interventions, and the proportion of c-units containing errors.

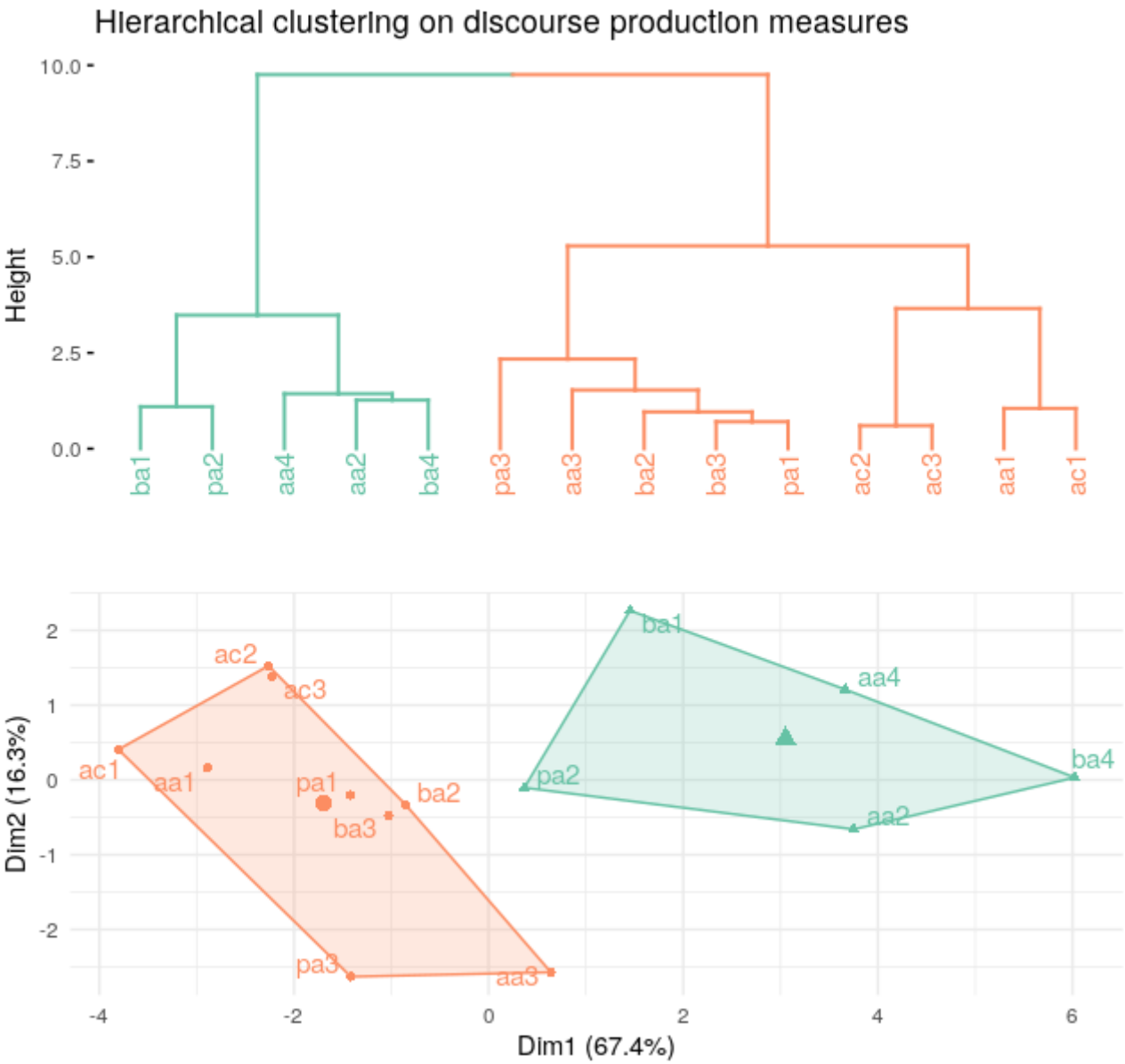

Figure 22: Cluster dendrogram of a hierarchical cluster analysis performed on selected discourse production measures. The lower panel shows the distances between individual participants in a two dimensional space; percentages show the amount of variation accounted for by the dimension.

We see that all non-fluent participants are grouped in one cluster and the fluent speakers with aphasia are clustered together with neurotypical participants. This suggests that the selected measures can reliably identify speakers with non-fluent aphasia. The inner structure of the respective clusters suggests that participants aa2, aa4, and ba4 share a more similar profile within the non-fluent group. Similarly, we see that the neurotypical speakers are grouped together with aa1. The inspection of the two dimensional projection brings additional information in the picture. First, the x axis can be interpreted as severity with the positive values indicating greater severity. The interpretation of the y axis is less clear, however, judging from the position of individual participants in the space and their language profiles, the y axis is most likely connected to productivity, particularly lexical diversity. Second, these groupings, based on the combined measures point at several relatively distinct language profiles within the sample.

Starting with the non-fluent speakers, we have seen that ba4 has the most severe aphasia from the whole group based on the majority of the measures. His speech is characterized by single word utterances or listings of nouns with very few verbs. Rather than narrating a story, ba4's discourse is therefore characterized by enumerating objects and participants in the story. There is a large number of phonemic paraphasias connected to the participant's having oral apraxia and also a high proportion of anomic pauses.

Participants aa2 and aa4 are close to this profile but their aphasia is less severe. Both participants form very short c-units that typically do contain a predicate, although the average number of verbs is around 0.5. A large proportion of c-units with predicates are variations of the Czech existential-presentative construction of the type 'there be X' (cf. also section 4.3.1). Their speech is characterized by lower lexical diversity and a great number of anomic pauses and abandoned c-units.

Both pa2 and ba1 are farther from the "core" non-fluent participants. The case of pa2 is interesting in that according to most of the measures he performed similarly to some of the less severely impaired fluent participants, as evidenced by the relatively short distance from participants pa1, ba2, and ba3. Note also that the position of pa2 in the two dimensional space is close to [0;0], suggesting that the participant is, as it were, an intermediary case, being not as severely impaired as aa2, aa4, and ba4, but also not as fluent as the speakers with mild-to-residual aphasia and the neurotypical participants. Subjectively, I assessed the linguistic behavior of the participant during the interview as highly non-fluent, but this is mainly reflected in the number of interventions and prompts on my part, which, as already mentioned, was caused by a marked slow-down in processing and a very hesitant production profile. This characteristic is also reflected in a low total number of c-units. A somewhat lower lexical diversity and a lower proportion of story relevant vocabulary reflect problems of noun retrieval. Lastly, ba1 is very far from the remaining participants. Her production is characterized by a smaller number of total c-units produced and a relatively high rate of disfluent chunks, resulting in a low number of fluent trigrams. These measures as well as a lower proportion of function words and a higher ratio of nouns to nouns and verbs place the participant in the non-fluent group. On the other hand, she produced more multiclause units than the other non-fluent participants and her speech has a higher lexical diversity.

When we look at the other cluster, we see participants aa3 and pa3 as one subgroup within the fluent speakers. aa3's severity is slightly higher, but other than that both participants share a distinct language profile discussed throughout this section. Their speech is characterized by relatively long c-units that contain a high number of highly frequent lemmas and have a low lexical diversity. They use a high number of pronouns and discourse markers which is reflected in a lower proportion of content words as well as a low ratio of nouns to nouns and verbs. We also find some discourse level errors in aa3 and pa3's speech, caused by the use of pronominal reference without clear antecedents. Compared to MLU, these participants produced relatively fewer fluent trigrams. All of this suggests a relatively severe impairment of lexical retrieval of nouns, as also evident from a high frequency of retracing and administrator interventions.

ba2, ba3, and pa1 form a group that is fluent and whose speech is close to the neurotypical speakers. What sets these participants apart from the neurotypical group is mostly fluency. These three speakers produced fewer fluent trigrams and their c-units contain more disfluent chunks. Their lexical di-

versity is also lower compared to ac1, ac2, ac3, and aa1 and about 15 % of their respective c-units contain errors. While this number as such is relatively low, it is substantially higher than in the most fluent subgroup. Lastly, aa1 is grouped with the neurotypical speakers which is consistent with my subjective assessment. The participant frequently had word finding difficulties during the session, but his speech was relatively fluent with few reformulations or abandoned utterances and rich in information.

In conclusion, I hope to have provided a first characteristic of connected speech production in Czech speakers with aphasia. This characteristic has found patterns consistent with those described in the literature for other, typologically different languages which is in accordance with Prins and Bastiaanse (2004)'s review of discourse analysis in aphasia in which they claim that existing body of research has found no major language specific difference at this level of language production. The measures used in the present work were derived from the most frequently used micro-structural discourse indices in the literature and most of the measures successfully differentiated between non-fluent aphasia and neurotypical speech. The failure to detect reliable differences between speakers with fluent aphasia and neurotypical participants was, by and large, caused by the specific profile of participants aa3 and pa3. Further analyses of larger samples of speakers are needed to establish if the proposed measures could also differentiate between neurotypical speech and anomic and residual aphasia. I also believe that the measures could be used in clinical practice with relative ease as they do not require advanced knowledge of linguistic theory and their coding is relatively straightforward and does not necessarily require a detailed token level analysis. The remaining three sections are dedicated to the demonstration of how the usage-based perspective can reveal patterns of linguistics behavior that may have implications for language assessment and therapy in aphasia.

4.3 Verbs and nouns in the corpus

Having provided a detailed characteristic of the participant's discourse production, I move on in this section to describe the uses of verbs and nouns in the corpus with special attention to frequency characteristics. In the first part I describe the production of verbs in the subcorpus with a focus on both grammatical and syntactic factors as well as cumulative lemma frequency. The analysis of nouns is concerned with possible effects of the relative frequency of individual word forms and paradigmatic cells of inflected nouns.

4.3.1 Lexical verbs in the subcorpus

The study of verb production and processing has been one of the focal points in the linguistic study of aphasia. Three main reasons can be formulated for this interest. Firstly, verbal predicates and their valency frames as well as the argument structure constructions they are correlated with are at the core of propositions and, consequently, clauses and utterances as the crucial buldings blocks of communicating the state of reality and the "inner world" of speakers and their beliefs and assessments of external states of affairs.[44] Secondly, it was already mentioned that the lingusitic research of aphasia has been, from a theoretical and methodological perspective, dominated by structuralist, rule-based approaches represented mainly by generative grammar. Generative grammar has been, in turn, mainly interested in syntactic structure and a point can be made that the syntactic structure of a given clause is by and large driven by the predicate and the configuration of the verb's arguments with regards to constituents and grammatical relations. For instance a passive clause is "generated" based on the fact that a theme argument of a transitive verb is to be used in the subject position of the given clause. This "syntactocentric" view drives the interest to understand how verbs and clause structure function in impaired language. The third reason comes from clinical practice and the observation that verbs can be markedly more affected compared to other parts of speech in some aphasia types. It has long been established that speakers with non-fluent aphasia, particularly with Broca's aphasia produce discourses with fewer verbs, have difficulties in confrontational verb naming, and encounter problems in verb comprehension (e.g. Rossi & Bastiaanse 2008; or Flanderková 2019). Numerous frameworks and theories have been proposed to explain these difficulties, such as the Argument Structure Complexity Hypothesis, the Trace Deletion Hypothesis, or the adaptation strategy proposals, discussed in section 2.2. Recall also that the few experimental studies conducted on Czech aphasia by Flanderková and collaborators have focused on verb processing (Flanderková et al. 2014; Hudousková et al. 2014).

The aim of this section is to describe verb production in the subcorpus and to explore any potential differences between participants with different aphasia types and language profiles. I focus on variables that are derived from argument structure, such as transitivity, number of arguments, or grammatical voice. These were selected because a body of research conducted has shown that speakers with agrammatic aphasia have more difficulties when processing verbs with more complex argument structure (Thompson 2003) and verbs that can be analyzed as involving movement, such as

44 The term argument structure construction is specific to Construction Grammar and refers to abstract clause schemas that are associated with the expression of certain types of events and actions (Goldberg 1995), e.g. the English ditransitive construction would have a preverbal subject slot expressing a "transferrer" and two postverbal slots with a NP expressing the transferred object and a to-PP expressing the "transferee". Such a generalized construction combines with sematically compatible verbs such as *give*.

passives or unaccusatives (Bastiaanse & van Zonneveld 2005). I was therefore interested to see if it was the case that some of the participants in the sample produced markedly fewer verbs falling within these categories, particularly the non-fluent participants aa2, aa4, and ba4 who presented with some traces of agrammatism during the interviews. Frequency of use is the other variable of interest in the present analysis. I already discussed in section 2.5 that one of the main contributions of linguistic analyses of aphasia employing the usage-based framework has been the finding that frequency-based constructional preferences of individual verbs modulate the processing of transitive and unaccusative uses of verbs (i.e. the Lexical Bias Hypothesis (Gahl 2002)). Such an analysis is outside the scope of the present work and a more simple and coarse grained assessment of the influence of cumulative lemma frequency on ease of retrieval is presented in this section. I explore not only the number and type of verbs produced by participants but also disfluencies that occurred in the context of verb retrieval in the data. The occurrence of disfluencies such as hesitation sounds or silent pauses was used as a marker of increased processing load during word finding (see section 4.4.2).

4.3.1.1 Data preparation

A total of 1291 tokens of lexical verbs, including phonemic paraphasias with clearly recoverable targets, were extracted from the subcorpus and further annotated to compare verb production of individual participants. The neurotypical speakers ac1, ac2, and ac3 were included in the analysis for reference. The verbs were annotated for the following variables.[45]

- semantic class: All of the verbs were grouped into semantic classes based on the basic Hallidayan distinction between material, mental, and referential process types (Halliday & Matthiessen 2014: chap. 5), behavioral and verbal types that also appear in Halliday's classification were grouped with the mental type.

- verb lemma: Reflexive pronouns were included as part of the lemma such that reflexive and non-reflexive uses were counted as two unique types, e.g. *vzbudit* 'wake someone' and *vzbudit.se* 'wake'.

- presence of any disfluency preceding the verb: All silent pauses, hesitations, word fragments and "intrusive" markers were coded as disfluencies. Cases in which the disfluency appears in the "auxiliary" complex of the lexical verb or before a reflexive pronoun immediately preceding the verb were also included (17.d and e). Repetitions were treated such that only the last repeated token was included in the analysis. Retracing was treated as follows. If the repair targeted a phonemic paraphasia or a grammatical error, only the repaired token was included. In cases where a produced verb was repaired to a different lemma, both tokens were kept for analysis and this information was marked in the coding scheme.

  17. Examples of different context coded as disfluent

     a. *a H 0.2 dělá takove baletní představení* 'and he's H 0.2 doing sort of a ballet performance' (ac2: 193); this represent the standard and most frequent situation when the verb is preceded by silent and/or filled pauses

45 Note that some of the variables were already available from the subcorpus, e.g. lemma frequency.

b. *a byli spolu šš̆asny* ‘and they were were happy together’ (pa3: 81); direct repetitions were counted as disfluencies

c. *a řekl bych že tím to víceméně skončilo když potom s- teda slezl dolů* ‘and I’d say that with that the story was more or less over after he c- well climbed down from the pole’ (ac1: 235); a small number of tokens contained a word fragment followed by a marker of repair, such as *teda* ‘well, actually’ or *nebo* ‘or’

d. *von H tam H se usmál a* ‘there H he H REFL smiled and’ (aa1: 71); disfluencies occurring before the sequence AUX/REFL + verb were included in the analysis

e. *takže tu logiku 0.5 H 0.2 sem nepochopil* ‘so the logic I 0.5 H 0.2 didn’t understand’ (aa3: 232)

- whether the verb is part of a multiverb construction: Verb tokens occurring in a modal or phase verb construction were coded as multiverb (e.g. *mám to vyprávět celý znova jako* ‘should I tell the whole of it again, then’) to account for additional syntactic complexity.

- voice: Verbs were classified as active, anticausative and passive, medium and reflexive, and reflexive tantum.

- number of obligatory arguments: The number of arguments was decided according to Vallex 4.0 (Lopatková et al. 2020). Verbs that lack an entry in this dictionary were annotated analogically to semantically related entries (e.g. *zabouchnout* ‘to slam shut’ was based on the frame of *zavřít* ‘close’).

- transitivity: Verbs were coded as transitive or intransitive based on the Vallex entries. Passive and anticausative uses were marked intransitive. Verbs that take either a direct object or a clausal complement were coded as transitive in the former situation and as having an intransitive frame in the latter.

- semantic roles in argument structure: A complete list of semantic roles of the given verb was coded based on its Vallex entry and the semantics of the verb. A simplified coding was developed with the roles summarized in Table 17.

role	characteristic
agent	agent. volitional argument with control over the event expressed by the verb
patient	patient or theme, stimulus
undergoer	non-volitional argument affected by the verb
recipient	recipient
spatial role	any spatial role (goal, location, source, etc.)
state or property	prototypically a copular complement in the construction X is Y
verb	used for lexical verbs that have a verbal complement (the verbs *začít/začínat* 'begin' and *přestat/přestávat* 'stop')

Table 17: Overview of semantic roles used in the annotation.

- presence and type of direct object in transitive verbs: Transitive verbs were annotated for direct object expression (NP, pronoun, clausal complement, no object), verbs that do not require a direct object were marked accordingly.
- Cumulative lemma frequency was extracted from spoken and written corpora and log-transformed, lemmas without corpus attestation were coded as zero frequency and excluded from relevant analyses.

Using this annotation, I proceed to compare verb use by individual participants and groups.

4.3.1.2 Description of data

The analysis of the data has two main objectives. First, I explore individual profiles and differences with regard to the morphosyntax of the verbs, i.e. the use of tense and voice and argument structure, to the predicate classes produced, and to cumulative lemma frequency. I subsequently focus on the potential relationship of these variables with the occurrence of disfluencies.

Table 18 summarizes the total number of verb tokens produced by individual participants with the total number of c-units produced over the three tasks in the subcorpus and the mean number of verbs produced per c-unit repeated here from Table 11 in Section 4.2.1. Before I proceed to describe the verb production patterns in more detail, I would like to reiterate here the basic observation that the non-fluent speakers produced less verbs than both the fluent and the neurotypical group which is in line with previous research.

participant	group	verb tokens produced	c-units produced	mean verb per c-unit
aa1	clinical	125	103	1.27
aa2	clinical	80	159	0.616

aa3	clinical	144	168	1.03
aa4	clinical	34	77	0.468
ba1	clinical	32	41	0.805
ba2	clinical	139	122	1.18
ba3	clinical	120	114	1.15
ba4	clinical	10	89	0.135
pa1	clinical	95	87	1.11
pa2	clinical	48	51	0.941
pa3	clinical	135	106	1.38
ac1	nbd	142	108	1.47
ac2	nbd	120	108	1.21
ac3	nbd	67	63	1.14

Table 18: Overview of verb production of individual participants.

4.3.1.3 Tense and voice

Figure 23 summarizes the use of the past and non-past verb forms produced by the participants. Notice that the proportion of past and non-past tense uses is plotted for the three tasks individually. The reason for this is that CHAPLIN and COMIC aimed to elicit narrative discourse while descriptive discourse was expected for LADA. This is reflected by the distribution of past and non-past forms across the tasks for the majority of participants, regardless of fluency. The only exceptions are ac2 and, to an extent, ac3 who preferred the use of non-past across the tasks. We see an increase in non-past uses in the COMIC task which may be explained by the fact that the elicitation of narrative discourse may have been less successful in this task compared to CHAPLIN. This may have been caused by the nature of the stimulus, i.e. three static pictures in a sequence and the relative difficulty of the task. The predominant proportion of non-past forms in LADA indicates a successful elicitation of descriptive discourse. This comparison also suggests that there are no preferences for the past or the non-past in the individual speakers of aphasia regardless of diagnosis. The pattern of use of ac2 and ac3 can be explained in terms of idiosyncratic narrative style differences.

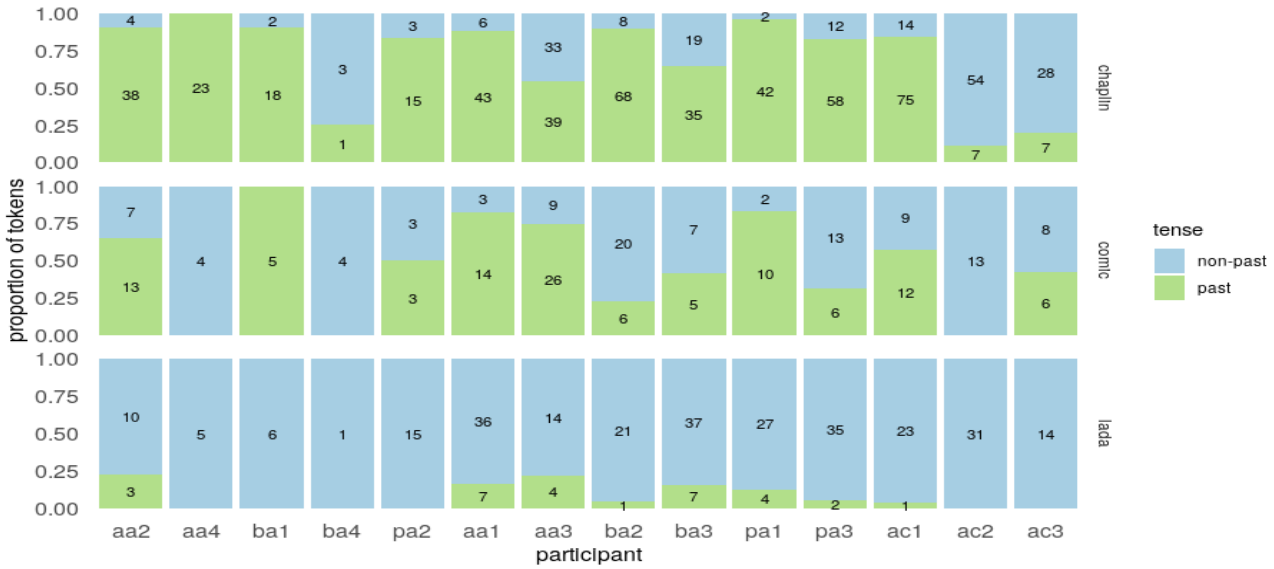

Figure 23: Proportion of past and non-past tense uses across the three tasks; number in column indicate token count.

However, two things do deserve a further comment. First, it should be kept in mind that the participants aa4 and ba4 produced a very small number of verb tokens, ba4 produced a total of nine finite verb tokens (4, 4, 1), while aa4 produced a relatively high number of tokens in the CHAPLIN task (23), but only a limited number of forms in the other two (four and five respectively). Secondly, an interesting case is the proportion of tenses produced by aa3 in the chaplin task. The proportion of both tenses is almost 1:1, which is very unusual compared to the other participants. The visualization of past and non-past uses across aa3's c-units in Figure 24 shows three "switches" between predominantly past and non-past verb turns, approximately around turns 70, 130, and 180. A closer inspection of the data shows that these switches were caused by sections of the narrative where the participant encountered greater problems with lexical retrieval and the present tense dominated parts correspond to collaborative repairs and commentative utterances of the speaker.

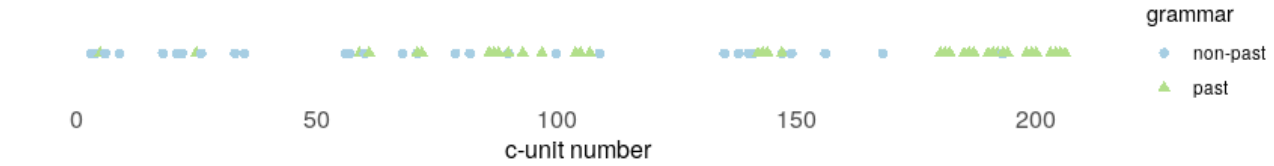

Figure 24: Production of past and non-past tense by participant aa3 in individual c-units in the chaplin task.

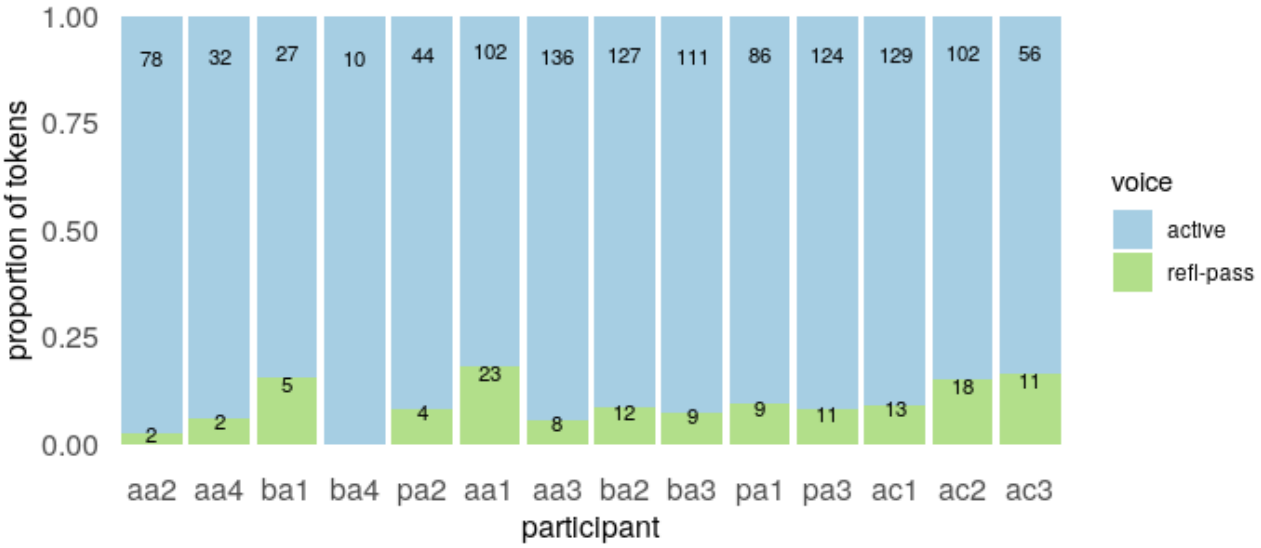

Figure 25: Proportion of active and passive-reflexive voice uses; numbers in columns indicate token count.

Figure 25 shows the proportion of the active and "passive-reflexive" verb uses. Note that tantum reflexive verbs were grouped with active non-reflexives, while anticausatives were grouped with reflexive and medium uses. Again, we see an expected pattern of a high proportion of active voice uses. A group comparison between fluent (participants with aphasia and neurotypical speakers grouped together) and nonfluent participants did not reveal any significant differences ($U = 25$, $p = 0.33$). A look at individual production suggests that differences in the use of the active and passive-reflexive voice are most likely driven by individual framings of the narratives and descriptions. Another contributing factor in participants aa2, aa3, pa3, and, to an extent, ba2 and ba3 is a lower lexical diversity and reliance on a lower number of highly frequent verbs, typically *be*, *have*, and *go*.

4.3.1.4 Transitivity and number of arguments

The next step was to focus on argument structure and transitivity as these have been reported to modulate verb production at least in agrammatic aphasia. When we look at transitivity first (Figure 26), we see that the use of transitive and intransitive verbs is relatively similar across participants with the non-fluent speakers aa2, aa4, and ba1 having even produced an above-average proportion of transitive verbs. The similar may be said about the number of arguments that are indicated as obligatory in their valency frames, as shown in Figure 27. Participants produced a majority of two- or three-argument verbs. Note that these counts refer to verb tokens rather than types and that they include intransitive verbs with two arguments. With one of the most frequently used verbs in the subcorpus being *být* 'be' it might be hypothesized that these proportions are due to a generally high frequency of the copular construction. However, this turns out not to be the case. When this verb is excluded the pattern stays the same.

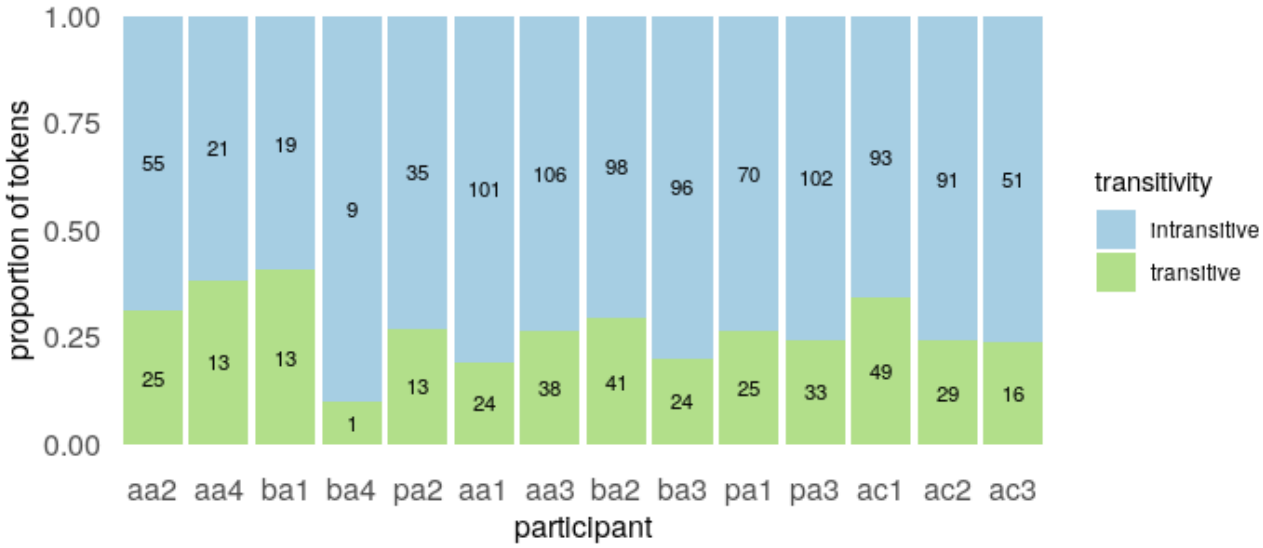

Figure 26: Proportion of transitive and intransitive verbs; numbers in columns indicate token count.

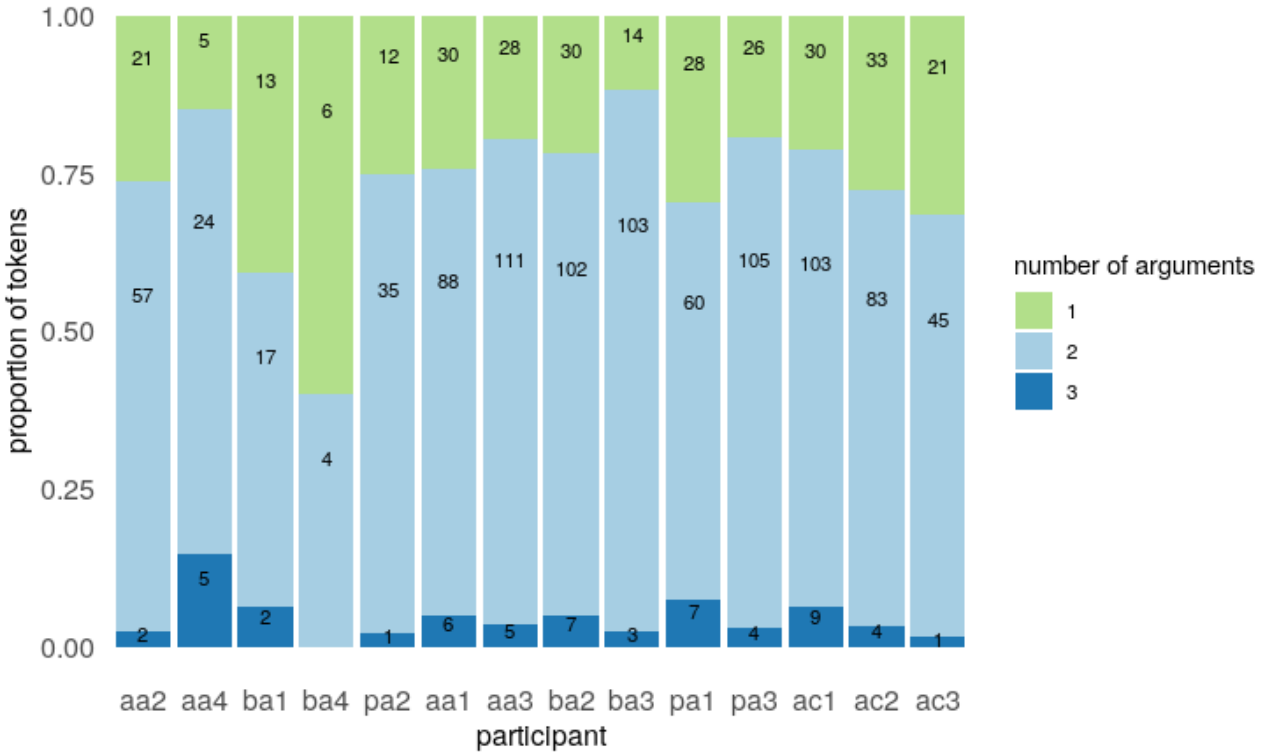

Figure 27: The proportion of uses of verbs with one, two, and three obligatory arguments; numbers in columns indicate token counts.

In order to account for this possibility that it is not just transitivity and number of arguments but rather the syntactic complexity of the argument structure that influences verb production, I created an additional variable, argument complexity, based on the hypotheses of Thompson (2003) and Bastiaanse and van Zonneveld (2005). Verbs were coded as being "argument-structurally" complex if it was used transitively, if it had the "passive-reflexive" voice label, or, to account of unaccusativity, if

the semantic roles in the verb's argument structure included only a patient or an undergoer, defined for the purposes of the present work as a (potentially) sentient referent that has something happen to themself that the undergoer does not have any influence over. Figure 28 shows the proportion of uses of simplex and complex verbs across participants.

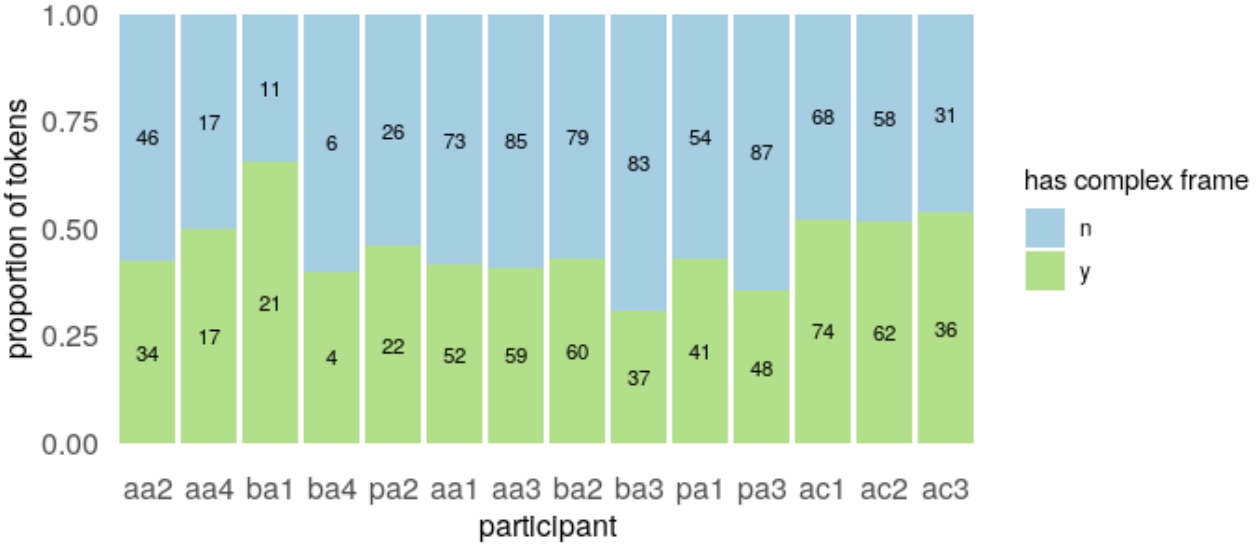

Figure 28: Proportion of verbs with complex and simplex argument structure; columns indicate token counts.

We see that participants aa4 and ba1 have a relatively high proportion of complex frames. This is due to a relatively high number of transitive verbs produced by aa4, while ba1 also produced a number of unaccusative verbs. However, in group comparison, the group of neurotypical speakers has a slightly higher statistically significant ($U = 3$, $p = 0.0385$) overall proportion of uses of verbs with complex frames. This difference is relatively small; the proportion of complex argument structure uses was above 0.5 in all three neurotypical speakers, while the values of participants with aphasia were on the whole below 0.45 (mean = 0.436 (sd 0.089), median = 0.425 (interquartile range 0.04)). Easier access to such verbs and constructions in non-brain-damaged speakers thus cannot be ruled out. However, one should be careful to interpret this difference too strongly given the size of the sample and further research would be needed.

An additional parameter that I explored in relation to argument structure and transitivity was the proportion of transitive verbs that lack an overtly expressed object in the data. As can be seen in Figure 29, this proportion is relatively high in participants aa2, aa3, and aa4. While some of the c-unit produced with no object are effectively ungrammatical others are just cases of verbs with a contextually or lexically retrievable object or class of objects that are perfectly acceptable in Czech (cf. 18.a and b). However, it is important to keep in mind that the production of a verb with a direct object drop, such as 18.b is still an active choice on part of the speaker who could have opted for an utterance such as *maminka něco vaří* 'mom is cooking something' and decided not to, based on their aims and communicative context.[46] Thus, rather than interpreting this trend visible in some of the participants as having to do with a reduced capability to process predicates with more complex

46 It also follows from this discussion that a majority of participants did opt for a strategy with an explicit object in similar contexts.

valency structure, the increased frequency of predicates used without direct objects may be a form of adaptation strategy to reduce risk of cognitive overload in the sense of Kolk and colleagues (e.g. Kolk & Heeschen 1992; Kolk 1995; Hartsuiker & Kolk 1998).

18. Transitive verbs produced without direct object

    a. *maminka 1 hledá 0.2 H 0.5 H hledá 3 H* ‘mom 1 is looking for 0.2 H 0.5 H looking for 3 H’ (aa2: 245); a similar utterance would only be acceptable if the object was already mentioned and highly activated from preceding discourse, which was not the case here

    b. *maminka 3 vaří* ‘mom 3 is cooking’ (aa2: 159); when focus is on the action of cooking as such, an object drop is perfectly acceptable in Czech

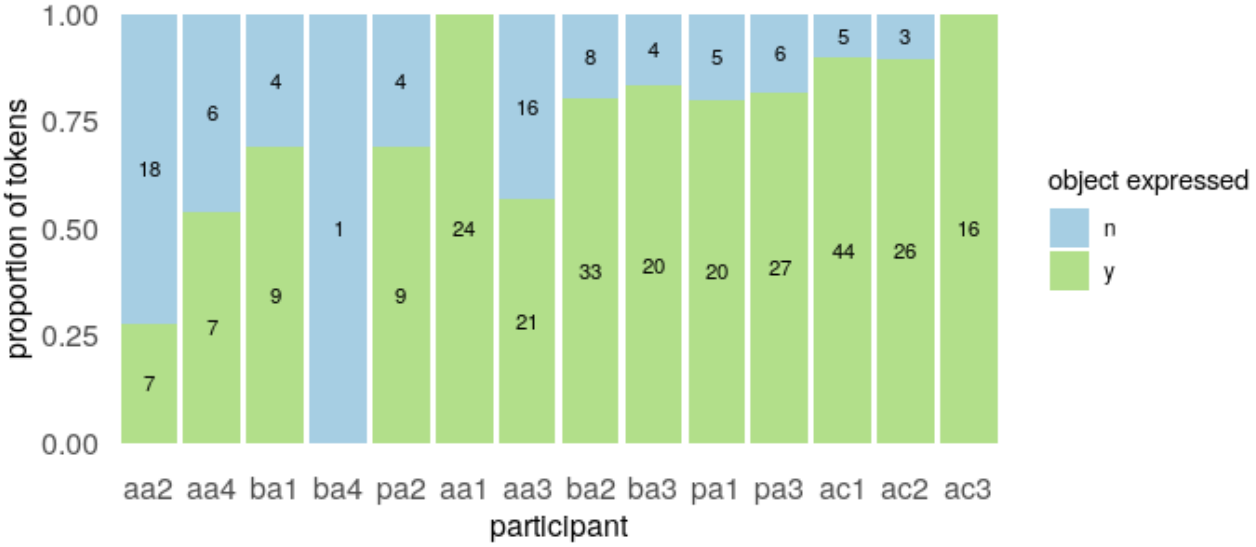

Figure 29: Proportion of transitive verbs with direct object omission; numbers in columns indicate token counts.

Figure 30 offers a summary of the proportions of Hallidayan process types used by participants. These are again plotted individually for each task because genre-based differences may be again expected. This is exactly what we find when we compare lada to chaplin and comic: there is a higher proportion of relational predicates which are, by and large, instances of the copular construction.

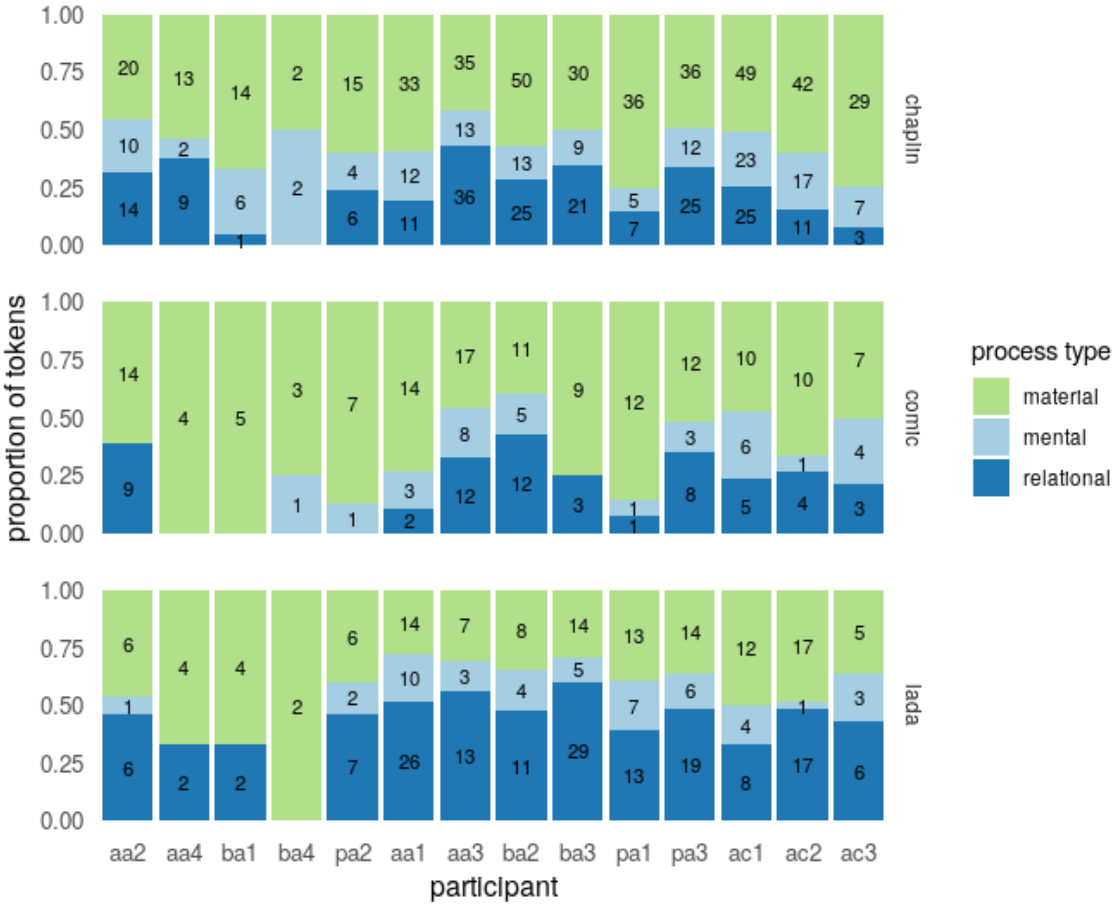

Figure 30: Proportion of verb process types across the three tasks; numbers in columns indicate token counts.

When we look at individual differences, we see a relatively higher proportion of relational predicates in the CHAPLIN task in participants aa2, aa3, aa4, ba3, and pa3. This high proportion of the referential predicates indicates that these participants produced a discourse with less narrative character. Moreover, participants aa2 and aa4 have a nonfluent aphasia with relatively higher severity compared to the whole group which is connected to an increased difficulty of verb retrieval and it is thus not surprising that a high proportion of their tokens are instantiations of the copular construction which has a very high frequency. In fact, four out of the 14 relational tokens and seven out of nine tokens, produced by aa2 and aa4 respectively, are instances of the existential-presentative construction with two word order variant *N tam byl* and *tam byl N* 'there was a N' as exemplified in 19.

19. *tam byl 0.5 H 0.5 židl- ne 0.5 ži- H 3 H 0.2 židle 0.5 ne 5 voda* 'there was 0.5 H 0.5 chai- no 0.5 chai- H 3 H 0.2 chair 0.5 no 5 water' (aa2: 82); note also that the verb form here is masculine while both the produced nouns are feminine which might be interpreted such that the participant relies on this partially fixed pattern *tam byl N* in situations of increased utterance planning difficulty

A similar point may be made for aa3 and pa3. The language profiles of these participants can be characterized with a high number of relatively long c-units which are subjectively relatively low in information value due to a high frequency of pronouns and discourse markers and a relatively lower

number of lexically specific items. An overuse of the copula construction fits well within this pattern. Recall also that Boye and Harder (2012)'s theory predicts that speakers with fluent aphasia will have problems with the retrieval of lexical items. While the copula in Czech would by their criteria be classified as lexical, one might argue that on the cline view of lexical and grammatical items copulas would be somewhere in a transitional zone between prototypical lexical and prototypical grammatical items. A closer look at participant ba3, who does not fit in either of the profiles described, reveals that nine out of the 21 uses of *být* occur in commentative or retracing "meta"-contexts, such as 20, which are not part of the main narrative/discourse frame.

20. Commentative uses of be in ba3's production (ba3: 26, 27, 29, 30)

    a. *no to to to není maringotka / to je něco něco jinýho no / to je něco jinýho / to (je) prostě klec 0.2 klec* 'well that that that isn't a caravan / that is something something else / that is something else / that is simply a cage 0.2 a cage

4.3.1.5 Frequency and lexical diversity of verbs

The last two factors to be discussed are frequency and type:token ratio. Figure 31 shows the distribution of log-transformed frequency of the verb lemma types produced by individual participants in oral. Reflexive and non-reflexive forms of verbs are treated as instances of the same type in this figure because these two uses are not differentiated in the corpus.

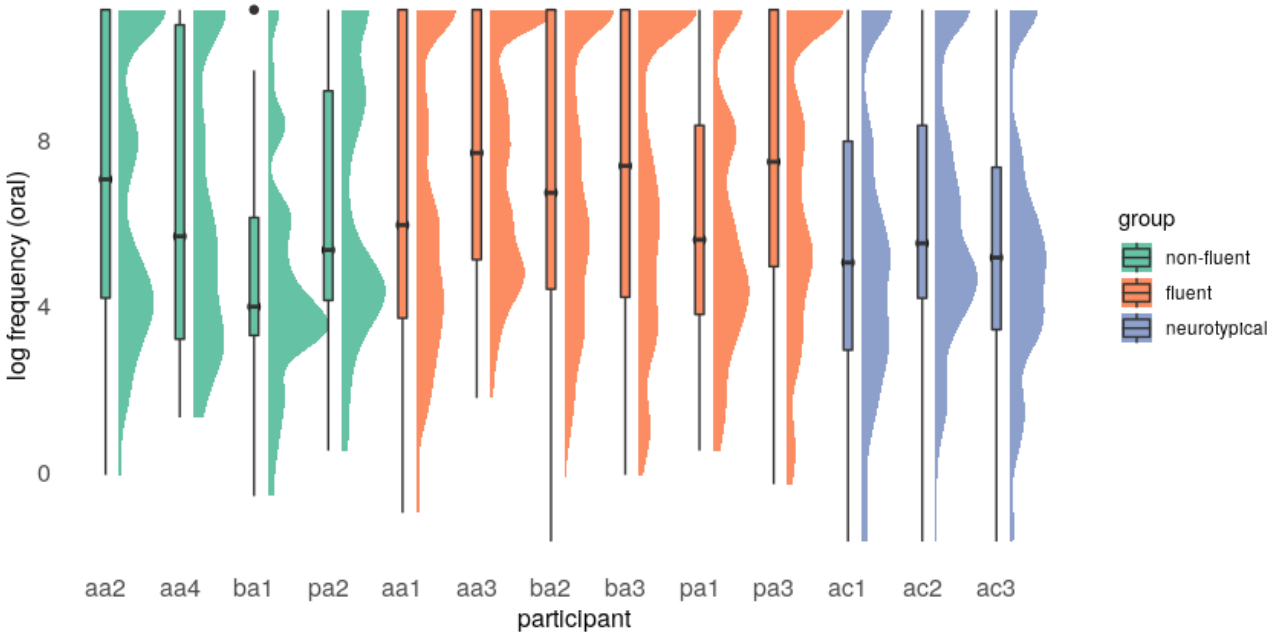

Figure 31: Raincloud plot of the distribution of log-transformed verb lemma frequency.

A comparison of the median values across the three groups using the Kruskal-Wallis test has shown a significant difference between the groups ($\chi^2_{(2)}$ = 6.79, p = 0.03), however a post hoc pairwise comparison using Dunn's test was not significant for any of the pairs. However, a comparison of individual profiles suggests that the fluent group produced verbs that are generally higher in frequency compared to both the nonfluent participants and the neurotypical speakers. When one looks

at individual verbs, it is apparent that particularly participants participants aa3, ba2, ba3, and pa3 produced a high number of tokens of the verb *být* 'be'. This result is not surprising given the fact that fluent aphasia types are characterized by lexical retrieval difficulties. We also know that frequency facilitates retrieval. The expected pattern for the group presented here would be such that there will be a lower number of verb tokens produced by the speakers with nonfluent aphasia, while the persons with fluent aphasia will produce more high frequency verbs, which is exactly the pattern that can be seen in the data.

This is further corroborated by lexical diversity of the verbs in the sample. As can be seen in Table 19 the type token ratio of participants aa3 and pa3 in particular are substantially lower and participants ba2 and ba3 reflect a similar trend. On the other hand, the nonfluent participants have very high TTRs which is explained by the fact that they produced a low number of tokens overall. The only exception to this is aa2 who produced a relatively high number of tokens and only a few types. This is related to a specific style of the participant, already discussed. First, rather than a narrative, he produced a descriptive discourse resulting in a high frequency of the existential-presentative construction. Second, there was a high number of persevering in the participant's production combined with word finding difficulties, as can be seen in 21. Notice that the pattern in a way overlaps with the existential construction (*there be N* v. *there run around N*). The high TTR of ac3 is caused by the narrative style of the participant who used a very concise and brief narrative in which the individual foregrounded component events of the story were stringed together without further elaboration.

21. Example of a persevering use of the verb 'run' in aa2's production (aa2: 17-19):

    a. *H 0.2 a 1 tam 1 běal chaplin 1 no 0.5 / a 0.2 tam 0.2 běhal ne 0.2 H 0.2 aha 0.2 / a 0.5 <sigh> 1 H 0.2 tam 1 běhal 1 H 1 H 0.2 tam běhal 1 chaplin 0.2 v kleci* 'H 0.2 and 1 there 1 was running around chaplin 1 well 0.5 / and 0.2 there 0.2 he was running around no 0.2 H 0.2 aha 0.2 / and 0.5 <sigh> 1 H 0.2 there 1 was running around 1 H 1 H 0.2 there was running around 1 chaplin 0.2 in a cage'

participant	token fq	type fq	ttr
aa1	56	37	0.661
aa2	44	19	0.432
aa3	84	33	0.393
aa4	24	15	0.625
ba1	21	17	0.81
ba2	88	47	0.534
ba3	60	32	0.533
ba4	4	4	1

pa1	48	31	0.646
pa2	25	19	0.76
pa3	73	25	0.342
ac1	97	61	0.629
ac2	70	47	0.671
ac3	39	33	0.846

Table 19: Lexical diversity of verbs produced by participants.

There is one more interesting pattern related to frequency and the way it interacts with the structural variables. While it has been discussed in the literature that patients with nonfluent, particularly agrammatic, aphasia have more difficulties in the production of verbs that are transitive and have more arguments in their valency frame, which is not directly reflected in this data. There is a tentative explanation for this. A closer inspection of transitivity and number of arguments with respect to frequency revealed a visible tendency for the high frequency verb tokens to be either transitive or intransitive 2-argument verbs. When we look at the verb tokens produced by the nonfluent participants, there is a significant weak correlation between lemma frequency and number of arguments ($r = 0.28$, $p = 0.00008$, 95 % CI = 0.144 – 0.404). It may thus be the case that the non-fluent participants do not produce comparatively fewer transitive (and "multi-argument") verbs overall because the transitive and thus, on some accounts, structurally more complex verbs produced are simultaneously high frequency words.

4.3.1.6 Disfluencies in verb production

In the remainder of this section I present an analysis of the relationship of verbs produced with disfluencies with the characteristics of the verbs. Disfluencies are treated as a proxy to difficulty of retrieval. The usage-based analysis of language in aphasia adopts the premise that problems in language processing are caused by a lack of processing resources/processing overload while the representations remain intact. "Mainstream" accounts of the processing impairments in aphasia are based on syntactic complexity. This predicts that verb tokens that are more complex because their argument structure requires more complements to produce a grammatical clause or as a result of a valency changing operation involving movement (under a generative analysis) will be more prone to error or disfluent production. On the other hand, the usage-based approach stresses the role of performance factors, particularly frequency, and would predict that verbs with low lemma frequency will be produced less fluently. In order to explore the contribution of these two clusters of variables to the fluency of verb production, I used the variables discussed above to explore their role fluency. Table 20 summarizes the number of fluently and disfluently produced verbs in the sample. The table

shows an expected pattern based on the language profiles of individual speakers. Participants with a higher proportion of disfluent chunks overall also produced more verbs disfluently.

participant	fluent tokens	disfluent tokens	proportion of disfluent tokens
aa1	108	17	0.14
aa2	37	43	0.54
aa3	104	40	0.28
aa4	19	15	0.44
ba1	15	17	0.53
ba2	77	62	0.45
ba3	80	40	0.33
ba4	3	7	0.7
pa1	83	12	0.13
pa2	37	11	0.23
pa3	113	22	0.16
ac1	121	21	0.15
ac2	96	24	0.2
ac3	61	6	0.09

Table 20: Number and proportion of verbs produced fluently and with disfluencies by participants.

Figure 32 show the proportion of fluent and disfluent production of verbs with more complex frames and unaccusative and unergative verbs. We see that both types of verbs have very similar patterns and no influence of argument structure complexity can be concluded.

Figure 32: Proportion of fluent and disfluent production in verbs with complex (upper panel) and simplex (lower panel) argument structure; columns indicate token counts.

When we turn to lemma frequency, a pattern begins to emerge. Figure 33 shows the distribution of fluent and disfluent production in relation to log-transformed ORAL frequencies. These figures reveal an interesting group pattern. We see a difference between the fluent and nonfluent participants such that the fluent participants have higher proportions of disfluent production in verbs with lower frequency. On the other hand, the participants from the nonfluent group have comparable proportions of fluency regardless of frequency.

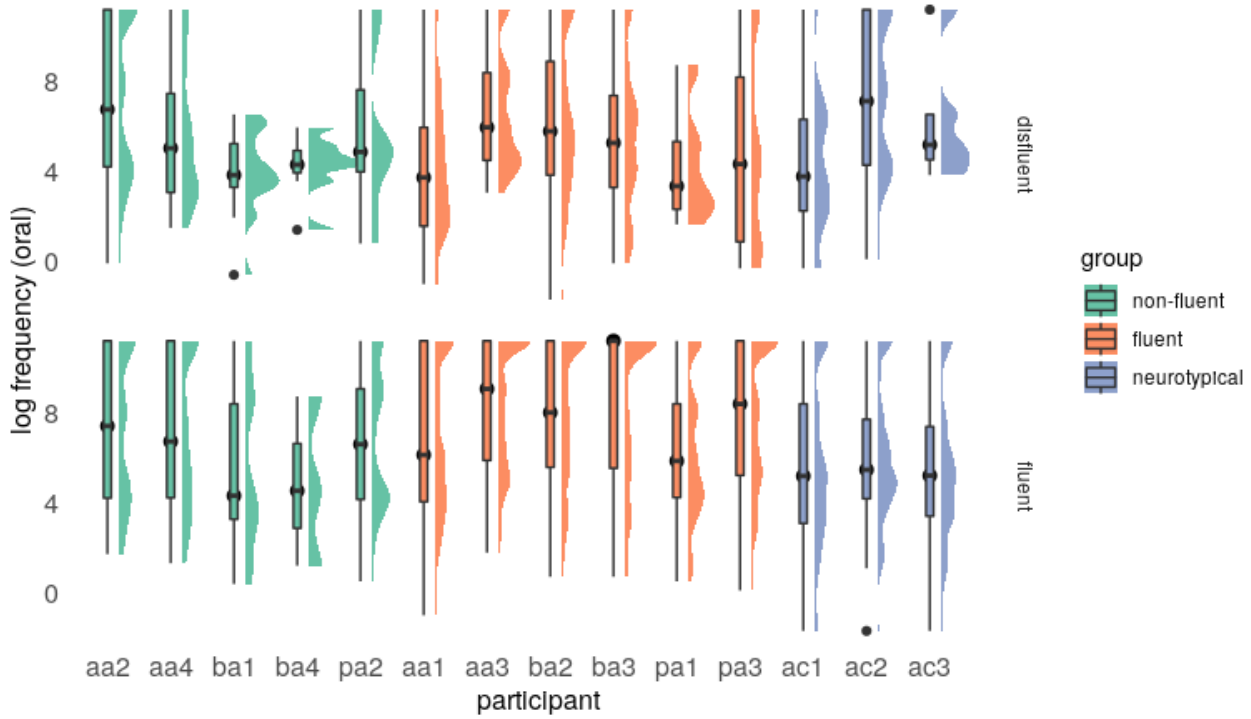

Figure 33: Raincloud plot of log-transformed lemma frequencies of verb tokens produced with disfluencies (upper panel) and tokens produced fluently (lower panel).

In order to explore this relationship further, I conducted a classification and regression tree analysis, using the following variables. Mean length of utterance, mean number of fluently produced trigrams, and mean number of disfluency chunks per word were used as a proxy for fluency. The log-transformed oral frequency of the lemma was used as a frequency measure and argument complexity was entered along with the annotation of multiverb construction occurrences to account for structural factors. I also included the Hallidayan process type as a combined characteristic of lexical semantics and valency. Two additional variables were included in the model: the position of the verb token in the c-unit and the total number of fully-formed words produced in the c-unit. The resulting model had the following structure: *disfluency occurrence ~ MLU + mean trigram count + difluency per word + oral log frequency + argument complexity + occurrence in multiverb construction + Hallidayan process type + position of token within c-unit + length of c-unit in words*. Participant ba4 was excluded from the analysis because he only produced ten verb tokens. The resulting solution is shown in Figure 34.

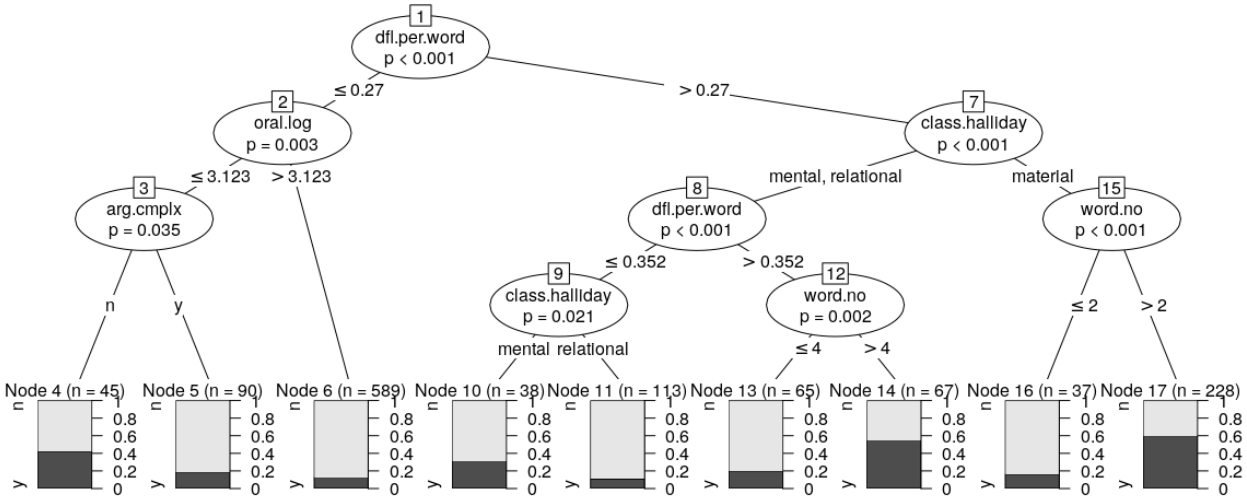

Figure 34: A classification tree of disfluent verb production.

We find the mean number of disfluencies per word as the most important factor in node 1. This is not very surprising given the fact that the dependent variable is the occurrence of disfluency. The nonfluent group delimited by the split in node 1 includes participants aa2, aa3, aa4, ba1, ba2, ba3, and pa3.

Focusing on the nonfluent group first, we see that the Hallidayan process type plays a role in node 7 differentiating between material verbs on the one hand and mental and relational verbs on the other. The split in material verbs in node 15 is based on the position of the token in the c-unit which shows that verbs that occur in the onset of the c-unit are on the whole more fluent, indicating that the farther the target word in the utterance is the higher the probability of disfluency is. It should be noted, however, that disfluencies and other para-linguistic phenomena were taken into account while numbering the sequences of tokens in c-units, i.e. first position of a verb in a c-unit necessarily implies its having been produced fluently. Seven out of 27 material verbs produced in position 2 are preceded by a disfluency in the sample out of the total 37 tokens in node 16. We see that material verbs produced by the nonfluent participants that are in position 3 or farther in the c-unit have on the whole a high proportion of disfluently produced tokens. No additional factor in the model was found to play a role in disfluency occurrence. Turning to the mental and process verbs under node 7, we find the number of disfluencies per word to play a role, differentiating between the more fluent participants aa3, ba3, and pa3 (node 9) and the less fluent participants (aa2, aa4, ba1, ba2). Node 12 supports the finding that position within c-unit plays an important role in this group with verbs occurring towards the onset of the c-unit. The split under node 9 differentiates between mental and relational verbs which are on the whole more fluent. The relational group is mostly comprised of the tokens of *být* ‘be’ and several tokens of *mít* ‘have’ which are both extremely frequent verbs, while the group of mental tokens has a high number of instances of *nevím* ‘I don’t know’ followed by *neumím* ‘I don’t know how to’ and *nepochopil* ‘(I) didn’t understand’. These are all highly formulaic expressions and account for a substantial number of fluently produced tokens.[47] However, one

47 It should be noted that particularly nevím occurs frequently as a discourse management device rather than a fully lexical verb. Such cases along with tokens such as *vidíte* ‘you see’ or *počkej* ‘wait’ which also function as pragmatic markers were excluded from the analysis and only “literal”, non-pragmatic uses of *nevím* were kept in the sample.

also finds cases of disfluently produced *nevím*, such as 22. However, these disfluencies can be interpreted as connected to a failed retrieval of a target word which is signaled by this token.

22. *no 0.2 nevím hm* 'well 0.2 I don't know hm' (ba3: 171)

If we focus on the more fluent participants, we see a clear effect of lemma frequency under node 2 with verbs with cumulative log-frequency over 3.123 (i.e. approximately 23 pmw) resulting in more fluent production. Finally, node 3 shows an effect of argument complexity which, however, suggests that verbs with simplex frames were produced with more disfluencies than the ones with complex frames. This simplex group under node 4 consists of tokens produced by the neurotypical participants and by aa1 and pa1.[48] On closer inspection, a number of the disfluently produced tokens in this group includes reformulations and repairs, such as 23.

23. Representative examples of disfluent prodution of verbs with simplex frames

a. *no 0.5 pak přiběhla 0.5 přiběhlo děvče …* 'well 0.5 then came-f 0.5 came-n a girl-n …' (pa1: 27); the disfluency is not caused by the retrieval of the verb as such but rather by an agreement error caused by a mismatch between biological and grammatical gender of the subject noun (cf. example 14.c on page 84)

b. *ch- doch- přichází tam 0.5 ženská s H 1 nádobou …* 'g- arri- is coming there 0.5 a woman with H 1 a container …' (aa1: 152); the participant has activated several competing verbs and is trying to select the appropriate for the context (the word fragments suggest the verbs *chodit* 'go-IMPF' and *dojít* 'arrive')

We can thus see that rather than structural complexity, it is frequency (and fixity) and other performance factors, such as the position of the target verb within a c-unit which explains some of the disfluencies preceding verbs in the sample.

4.3.1.7 Conclusion

In this section I described the production of verbs by individual participants with respect to structural complexity and frequency. This exploratory analysis has revealed no systematic group differences in tense use, individual differences most likely reflect personal narrative styles. When we look at verb production in participants with lower levels of fluency and frequency is also accounted for, an interesting pattern emerges. These participants do not on the whole prefer intransitive or one-argument verbs. However, there is a correlation between transitivity and number of arguments on the one hand and lemma frequency on the other such that this group produced transitive verbs with an overall higher frequency, underlining the importance of including this variable in analyses of aphasic (verb) production. These participants also tend to produce more transitive verbs with direct object elipsis which may be interpreted as an adaptation strategy to decrease processing load. We also see an overuse of the existential-presentative construction *X být tam* which is in line with the general assumptions of the usage-based framework as this construction may be predicted to be highly entrenched and easily accessible which in turn would mean that more processing resources may be allocated to noun retrieval. The data also show that neurotypical speakers produced more verbs with complex argument structure as defined by Thompson (2003) and Bastiaanse and van

48 A single token was contributed by pa2.

Zonneveld (2005). Participants aa3 and pa3 produced verbs with an overall higher lemma frequency and with low lexical diversity which is consistent with their general language profile as described in section 4.2.4.

The analysis of the distribution of disfluencies with regards to argument structure complexity and frequency did not indicate that verbs with complex argument frames induce more difluencies. On the other hand, lemma frequency does predict the occurrence of disfluencies in more fluent participants. The Hallidayan process type is a predictor in the non-fluent group such that referential predicates, mostly be and have, which are incidentally also highly frequent, induce less disfluencies. Mental predicates were also produced more fluently which can be explained by the use of tokens that were used in a commentative function outside the actual narratives and descriptions, such as *nevím* 'I don't know'.

In conclusion, the production of verbs in the subcorpus has shown effects of lemma frequency and the interplay of the distributional characteristics of verb lemmas with argument structure. Despite the fact that this was not the objective of this particular analysis, there is also some indication of the role of (partially) lexically specific constructions. These findings stress the importance of accounting for these factors which are an integral part of the usage-based approach.

4.3.2 Nominal inflection

The remainder of this section is concerned with a similar, albeit much shorter, description of noun inflection in the subcorpus. Czech is a morphologically complex, highly inflecting language which is one of the reasons it may be of interest for usage-based aphasia research. The research on aphasia in morphologically complex languages similar to Czech has reported a tendency for the use of uninflected/citation/unmarked word forms (Lehečková 1986). The grammar of Czech noun inflection is of a substantial complexity. Nouns are inflected for gender, number, and case. There are three genders with several inflectional classes and case syncretism (grammatical homonymy) is widespread as evidence by #table with an overview of two feminine inflection classes. Importantly, particular word forms and paradigm cells do not have equal probability of occurrence in texts. A substantial proportion of occurrences of individual lemmas is typically accounted for just by two to four word forms/number-case combinations. This is to a large extent driven by lexical semantics of a given lemma, as evidenced in Table 21 which lists the first four most frequent number-case uses in a corpus of spoken Czech for the lemmas *knedlík* 'dumpling', *vlak* 'train', *byt* 'apartment', and *pes* 'dog'.

Based on the functions and meanings of cases, one can assume that an apartment is something one has (accusative) or lives/stays in (locative), a dumpling is something one makes or eats in greater quantities (plural accusative), a dog as a living creature appears frequently as an agent in events (nominative) or is something people own (accusative), and train is something people travel on (instrumental) or wait for (accusative).

byt			knedlík			pes			vlak		
num	case	rel fq	num	case	rel fq	num	case	rel fq	num	case	rel fq

sg	acc	.605	pl	acc	.321	sg	nom	.369	sg	ins	.348
sg	loc	.126	pl	nom	.183	sg	acc	.216	sg	acc	.216
sg	gen	.1	sg	acc	.172	pl	ins	.097	sg	loc	.111
sg	nom	.073	sg	nom	.086	pl	nom	.081	sg	gen	.097

Table 21: Grammatical profiles of the lemmas byt 'apartment', knedlík 'dumpling', pes 'dog', and vlak 'train' expressed as relative frequency of paradigmatic cells; num = number, rel fq = relative frequency.

This nature of Czech may be of particular interest to a usage-based study of aphasia, since some of the nouns have a high frequency of use in number-case combinations that differ morphologically from the citation form which may cause a tension between morphological complexity and frequency. The aim of this analysis was to explore the production of inflected forms by individual participants and their distributional characteristics. A very general expectation may be formulated based on the usage-based assumptions. One might predict that participants with reduced levels of fluency will produce a lower number of inflected forms and that they will produce a lower number of inflected items with low relative frequency. In a way, this analysis follows up on the findings of Lehečková (1986; 2001; 2009) who describes that speakers with agrammatic and paragrammatic symptoms tend to use the nominative, i.e. the citation form, instead of other, context appropriate case form. The present description is to the best of my knowledge the first to explore the relationship between relative frequency and the use of nominal inflectional morphology.

4.3.3 Data preparation

I extracted all 1181 fully formed noun tokens from the subcorpus and annotated those for the following variables. Both the speakers with aphasia and the neurotypical participants were included in this analysis.

- Gender, number and case of the token.
- Whether the produced word form was phonemically paraphasic.
- Whether the noun is a proper or a generic noun.
- Whether the word form differs from the citation form.
- Relative frequency: Using the lemmas, I consequently extracted full grammatical profiles, i.e. all noun-case combinations and their relative frequencies, for the nouns, using frequency data from ORAL and SYN2020. Rank order and relative frequency of the paradigmatic cell were computed for noun tokens.

Using these variables, I compared inflected and uninflected word forms produced by individual participants with regards frequency.

4.3.3.1 Description of data

Table 22 summarizes the token frequencies of generic and proper nouns produced by individual participants as well as the mean number of nouns per c-unit in relation to the total number of fully-formed words, number of c-units, and MLU (cf. Table 11).

participant	group.clin	generic nouns	proper nouns	mean nouns / c-unit	mlu	tokens	c-unit count
aa1	fluent	115	12	1.23	7.515	790	103
aa2	nfluent	116	2	0.74	2.830	508	159
aa3	fluent	84	4	0.52	5.381	957	168
aa4	nfluent	66	0	0.86	2.714	225	77
ac1	nbd	130	4	1.24	8.065	915	108
ac2	nbd	108	5	1.05	5.611	621	108
ac3	nbd	73	4	1.22	5.571	360	63
ba1	nfluent	47	0	1.15	3.780	157	41
ba2	fluent	105	6	0.91	5.254	663	122
ba3	fluent	99	3	0.89	5.991	759	114
ba4	nfluent	47	3	0.56	2.000	178	89
pa1	fluent	61	4	0.75	5.414	488	87
pa2	nfluent	35	1	0.71	4.235	221	51
pa3	fluent	45	2	0.44	7.057	806	106

Table 22: Production of generic and proper nouns by participants

We see that the numbers follow the general language profiles of the participants. Participants aa3 and pa3 stand out in this respect compared to the rest of the participants. These two participants share a profile characterized by relatively long utterances and a high number of tokens produced but a simultaneous low number of noun tokens. This is due to the high number of pronouns, adverbials, conjunctions, “particles”, and interjections, resulting in a pattern of relatively long utterances and discourses with low information load.

Table 23 shows lexical diversity of nouns across participants. Notice the difference between participants aa3 and pa3 who both produced a lower number of noun tokens while their TTR values are markedly different. One of the reasons for this is the clinical profile of aa3 who has conduction

aphasia. One of the hallmark symptoms of conduction aphasia is inability/impaired ability to repeat words or phrases. There are several places in the transcript of the aa3 session where the participant encounters lexical retrieval problems and, when prompted with the target word by the administrator, attempts to produce the word repeatedly which results in a series of phonemically paraphasic words (*conduite d'approche*). When these tokens are excluded from the TTR computation, the value rises to .559, which is still considerably lower than the TTR of pa3.

participant	type frequency	token frequency	type token ratio
aa1	74	127	0.583
aa2	50	118	0.424
aa3	39	88	0.443
aa4	26	66	0.394
ac1	76	134	0.567
ac2	79	113	0.699
ac3	53	77	0.688
ba1	27	47	0.574
ba2	62	111	0.559
ba3	62	102	0.608
ba4	31	50	0.620
pa1	42	65	0.646
pa2	28	36	0.778
pa3	35	47	0.745

Table 23: Lexical diversity of nouns produced by participants.

With the general properties of the nouns having been described, I focus on the use of inflected nouns by individual participants in the subcorpus. Only generic nouns were included in this part, for two reasons. First, only very few tokens were proper nouns and, second, 29 of a total of 50 proper noun tokens were instances of *Chaplin* and *Charlie* who the main character of the chaplin task and is almost exclusively used in the nominative (as the main "acting force"), while nine of the 16 inflected tokens are comprised by the name Josef Lada, the painter of the picture used in the lada task. While it is reasonable to assume that expressions such as *Josefa Lady* 'of/by Josef Lada' will be highly fixed and represented as chunks, the specific properties of proper nouns, particularly person names, make a comparison of corpus frequencies a little problematic. Given the small number of

mostly uninflected proper noun tokes, I decided that it would be better to exclude this group from the analysis without skewing in the data. While trying to produce the target noun *tygr* 'tiger', participant aa3 used spelling as a supporting technique. The individual phonemes were tagged as nouns in the corpus. However, corpus frequencies are not available for obvious reasons and these tokens were also excluded from the analysis. The final sample thus included 1126 noun tokens. The proportion of uses of inflected forms, defined here as word forms different from the citation, is summarized in Figure 35.

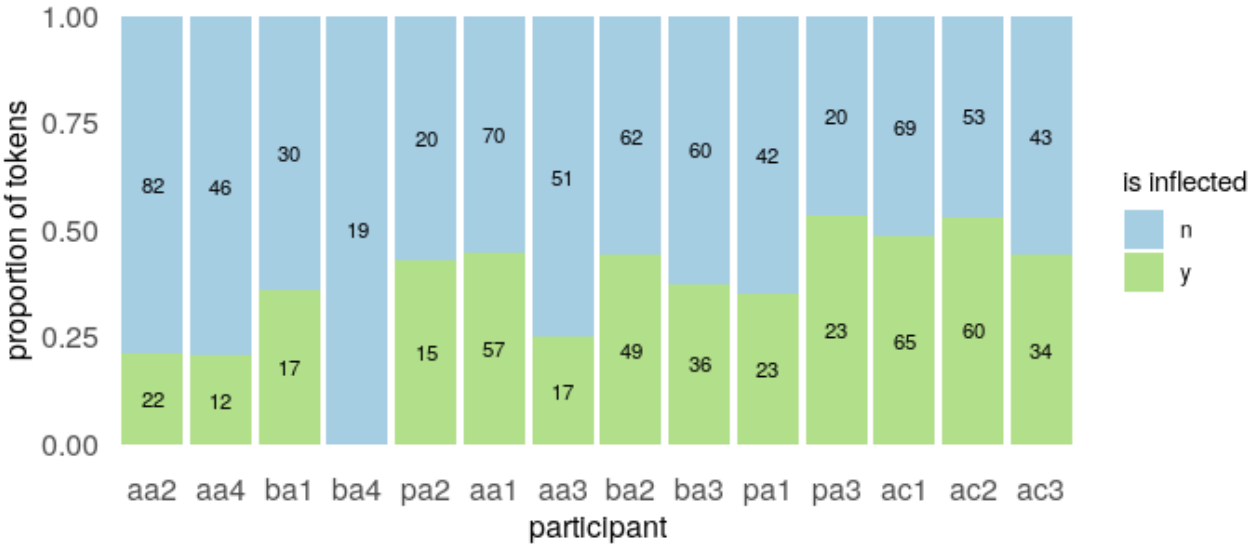

Figure 35: Proportion of inflected nouns; number in columns indicate token counts.

As can be clearly seen in Figure 35, participants aa2, aa3, and aa4 produced a markedly lower number of inflected tokens, while ba4 produced only a single token. The proportion in the remaining participants ranges from 0.33 to 0.53 (mean = 0.44 (sd 0.069), median = 0.46, inter-quartile range = 0.117). This pattern can be explored further by looking at the uninflected tokens with regards to case. Apart form the citation form, i.e. the nominative, the most frequent case with a form syncretic with the nominative is the accusative although some paradigms include more cells with uninflected forms, with the neuter class of nouns ending in *-í* represented mostly by verbal nouns such as *vaření* 'cooking' being the most extreme with only four cells different from the citation form. When we look at the distribution of uninflected nominative and accusative tokens, we find that the three participants with low inflection rates produced a substantially lower number of accusatives. This points to the fact that these three patients produced a higher number of nouns either in isolation or as subjects which in turn can be linked to an overuse of the existential-presentative construction. This suggest a qualitatively slightly different production pattern or strategy that can be said to be less demanding in terms of processing resources.

I subsequently explored the relationship of the inflected tokens with relative frequency. As noted above, high relative frequency of an inflected form is expected to make it more accessible and easier to retrieve. It is thus predicted that if the participants use an inflected form it will be more likely a high frequency one. While it might be a trivial thing to say, it is important to note that this prediction does not mean that speakers will generate inflected forms stochastically based on their mental

tallies rather than by the extra-linguistic context and the communicative needs and goals of a speaker. The scenario evoked by this general prediction may be conceptualized in two ways. Let us take the chaplin task as example. Say that a speaker is describing a scene where Charlie Chaplin sees a lion in a cage. A model sentence describing this might be something like *Chaplin viděl v kleci lva.* 'Chaplin saw a lion in the cage' with a PP with an inflected complement.

However, a person with aphasia may produce a less complex description, such as 24. Both the referents are expressed in independent existential-presentative constructions as subjects, i.e. nominative NPs. This construction may be assumed to be very frequent, especially in spoken Czech. A simple query of the spoken corpus shows that the combined frequency of the chunks *tam byl* 'there was' and *byl tam* 'was there' is 351 pmw. 24.a and b then shows that the information about the background of the scene, i.e. the cage, which is highly contextually available, may be omitted altogether without decreasing coherence and informativeness. 24.c is a hypothetical example of a further elaboration of the existential-presentative strategy which employs the same construction to communicate this backgrounded information.[49] These results should be approached with precisely this in mind.

24. Possible strategies for avoiding complex structures:

    a. *Chaplin tam byl a lev byl tam* 'Chaplin there was and lion was there'

    b. *Chaplin tam byl a lev (byl tam) v kleci* 'Chaplin there was and lion (was there) in cage'

    c. *Chaplin tam byl a lev byl tam a tam klec* 'Chaplin there was and lion was there and there cage'

All inflected tokens were used for the analysis, i.e. some word forms occurred multiple times and these were treated as individual usage events. It should be noted that in such cases when a given inflected word form occurs twice or multiple times within a single turn or in two or more adjacent turns, priming will be at play, facilitating the production of subsequent word forms. I decided to keep such cases in the sample. I also kept tokens that were produced with phonemic paraphasias that had clearly reconstructable targets. A total of 12 paraphasic tokens was included with mean edit distance to target 1.33 (median = 1). There was one nonword in the sample which was excluded as well as two tokens for which the word form and the context did not allow for unambiguous identification of number and case.[50] Tokens with no attestation in either corpus were treated as follows. Using the relative frequencies of the number-case combinations and word form, I also obtained rank frequencies of the tokens. The relative frequencies of word forms warrant one additional methodological remark. I wanted to account both for case-number combinations and word forms because

49 Note that these examples were created for the sake of the argument, but are based on the patterns observed in the corpus.

50 The nonword *štoudlík* had the target *štoudev* 'vat, cask' and was produced by ba3, it is possibly a blend of the target with another colloquial variant with similar meaning *štandlík*. While the target word was clearly identifiable, the neologistic token was excluded because of the large edit distance between the two and also based on the fact that štoudev is feminine in Czech, while the nonword štoudlík is unambiguously masculine due to its ending in *-lík* (cf. *balík* 'parcel' or *králík* 'rabbit'; all of the lemmas in syn2020 ending in *-lík* are masculine). The ambiguous word forms were two aa2's instances of *tyče* 'pole', one of these produced with a paraphasia as *čiče*. The context here was that the participant was describing Charlie Chaplin on a pole. The word form as such is syncretic between singular genitive and plural nominative and accusative. Since none of these fit well with the context, I decided to exclude these tokens.

Czech nominal morphology is highly syncretic and a single word form can have several different functions depending on the context. In addition to this fact, there is a substantial amount of variation in inflectional paradigms between standard (mostly written) and colloquial Czech, meaning that a single standard written form can have several variants in the spoken language. Since the spoken corpus understandably accounts for this variation in its transcription rules, which might skew the data, I decided to use the following procedure to obtain a good approximation of word form frequencies. The word forms in the sample were converted to their respective standard Czech orthographic variants and only the data from syn2020 was used for the analysis of word form frequencies, while both spoken and written frequencies were used for number-case combinations. A total of 425 inflected tokens were analyzed after these processing steps. Table 24 provides summaries of the relative frequencies for individual participants.

participant	number of inflected tokens	num-case rel. fq (oral)	num-case rel. fq (oral)	num-case rank fq (oral)	num-case rel. fq (syn2020)	num-case rel. fq (syn2020)	num-case rank fq (syn)	wf rel. fq (syn2020)	wf rel. fq (syn2020)	wf rank fq (syn2020)
		median (MAD)	mean (sd)	median (MAD)	median (MAD)	mean (sd)	median (MAD)	median (MAD)	mean (sd)	median (MAD)
aa1	54	0.111 (0.122)	0.143 (0.155)	4 (2.965)	0.169 (0.16)	0.177 (0.126)	2.5 (2.224)	0.262 (0.188)	0.251 (0.162)	2 (1.483)
aa2	27	0.186 (0.118)	0.159 (0.101)	3 (2.965)	0.158 (0.086)	0.158 (0.074)	2 (1.483)	0.259 (0.109)	0.233 (0.086)	2 (1.483)
aa3	15	0.232 (0.212)	0.331 (0.297)	2 (1.483)	0.261 (0.136)	0.28 (0.195)	2 (1.483)	0.334 (0.099)	0.38 (0.185)	1 (0)
aa4	14	0.158 (0.163)	0.163 (0.118)	2.5 (0.741)	0.15 (0.09)	0.117 (0.079)	3 (1.483)	0.15 (0.109)	0.133 (0.098)	3 (1.483)
ac1	63	0.156 (0.15)	0.212 (0.214)	3 (1.483)	0.137 (0.059)	0.172 (0.128)	3 (1.483)	0.225 (0.13)	0.261 (0.196)	2 (1.483)
ac2	57	0.137 (0.14)	0.173 (0.188)	4 (2.965)	0.145 (0.114)	0.183 (0.132)	2 (1.483)	0.236 (0.146)	0.247 (0.158)	2 (1.483)
ac3	32	0.168 (0.127)	0.194 (0.165)	3.5 (2.224)	0.147 (0.085)	0.191 (0.126)	3 (1.483)	0.224 (0.127)	0.263 (0.196)	2 (1.483)
ba1	17	0.227 (0.164)	0.278 (0.249)	2 (1.483)	0.202 (0.105)	0.21 (0.16)	2 (1.483)	0.208 (0.149)	0.289 (0.21)	2 (1.483)
ba2	49	0.14 (0.14)	0.183 (0.175)	3 (1.483)	0.137 (0.096)	0.167 (0.136)	3 (2.965)	0.229 (0.122)	0.241 (0.16)	2 (1.483)
ba3	36	0.156 (0.162)	0.198 (0.202)	3 (2.965)	0.14 (0.129)	0.181 (0.143)	3 (2.965)	0.295 (0.166)	0.275 (0.156)	2 (1.483)
pa1	20	0.158 (0.119)	0.157 (0.097)	3.5 (2.224)	0.179 (0.112)	0.18 (0.088)	2 (1.483)	0.223 (0.128)	0.213 (0.126)	2 (1.483)
pa2	16	0.101 (0.131)	0.173 (0.217)	3.5 (3.706)	0.122 (0.141)	0.16 (0.159)	4 (4.448)	0.155 (0.192)	0.246 (0.245)	3 (2.965)
pa3	24	0.226 (0.101)	0.222 (0.197)	2 (1.483)	0.166 (0.108)	0.188 (0.152)	3 (2.965)	0.209 (0.139)	0.226 (0.172)	3 (1.483)

Table 24: Mean and median values of the frequency of inflected nouns produced by participants.

Participant ba4 is excluded from Table 24 because he only produced one inflected token, *vetyny*, a paraphasic form of a target word form *květiny* ‘flowers’. This word form has a very low number-

case relative frequency in oral (0.067) which is very different from the syn data where it has a relative frequency of 0.167. A closer look at this difference provides a simple explanation. There are only 15 occurrences of the lemma in ORAL and four tokens of the word form *květiny* are tagged as singular genitive. However, a look at the contexts reveals that these are all mislabelings and the form has in fact the function of plural nominative or accusative. When one accounts for this fact, the values become much similar (relative frequency of 0.13 for plural nominative). In addition, a look at the word form shows that it is by far the most frequent word form of the lemma, accounting for over 45 % of all occurrences in SYN2020 and 53 % in oral. As a plural dominant lemma, it is not surprising that this is the word form that was produced by ba4 rather than the singular citation form *květina*.

When we compare individual participants we see a general pattern that goes in the predicted direction, i.e. all the participants produce number-case combinations and word forms that are frequent overall combined with a few items with a very high relative frequency and a varying number of items that are used relatively rarely in the produced word forms. This suggests that there are no qualitative differences in the sample and that all participants with aphasia present a similar behavior to the neurotypical speakers and individual differences are rather quantitative in nature such that individual participants produced different numbers of tokens from the particular relative frequency "bands". In other words, even in the speakers with comparatively lower proportion of inflections, we do not see participants who would only produce word form with very high relative frequencies that are (parts of) fixed, formulaic expressions, such as the forms *pořádku* 'order-sg.loc, *podstatě* 'nature, essence-SG.LOC,', or *chvíli* 'while-SG.ACC', the former two appearing almost exclusively in the PPs *v pořádku* 'in order, all right' and *v podstatě* 'basically, in fact', while the letter has a function of a temporal adverbial *(na) chvíli* '(for) a while'. However, some of the participants with aphasia did produce a limited number of infrequent word forms as discussed below. This general pattern is clearly visible in Figure 36 which shows the relative frequency of word form in SYN2020 for all the inflected tokens produced by each participant.

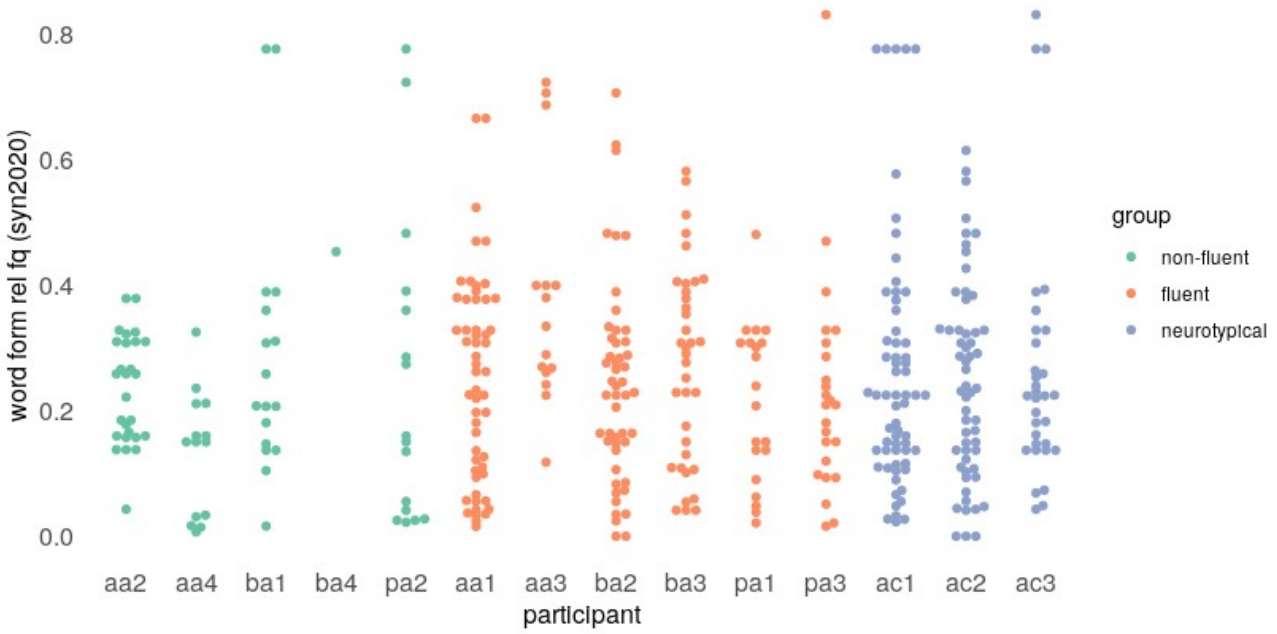

Figure 36: Relative frequency of word forms in syn2020 of inflected noun tokens.

When we compare the measures of central tendency for the number-case combinations in ORAL and SYN2020, we see some differences between the corpora. Individual differences are less extreme with regards to the SYN2020 frequencies. This can be partially explained by possible mistagging in ORAL (cf. the discussion of ba4's token *květiny*) and, primarily, by the fact that some of the expressions of the type *v podstatě* which are highly frequent appear on the whole are strongly associated with spoken language and, understandably, the broader range of topics, contexts, and constructions that appear in syn2020 increases the probability of a lemma's occurrence in certain word forms and number-case combinations. For instance, the word forms *zvířat* 'animal-PL.GEN' or džbánkem 'jug-SG.INS' have considerably higher relative frequency in SYN2020. This is driven by the way in which the written corpus is assembled as well as its size (ca. 100 M tokens v. 5 M tokens in ORAL). This clearly shows the importance of the selection of reference corpora or the need to consider different corpora in analyses within the usage-based framework.

We can now turn to the group identified above by a considerably lower rate of the production of inflected tokens, i.e. aa2, aa3, and aa4. A comparison of these three participants renders three distinct profiles. First, aa3's production is characterized by a high number of tokens with high relative frequency and a lower number of low frequency items. This holds both for oral and syn2020 data and for number-case and word form frequency. As can be seen in Figure 37, which shows the number of tokens with a relative word form frequency below 5 % produced by participants, aa3 produced no such token. The word form with the lowest relative frequency in aa3's data is znaku 'sign, logo, coat of arm-sg.loc' with a relative frequency of 12 %.

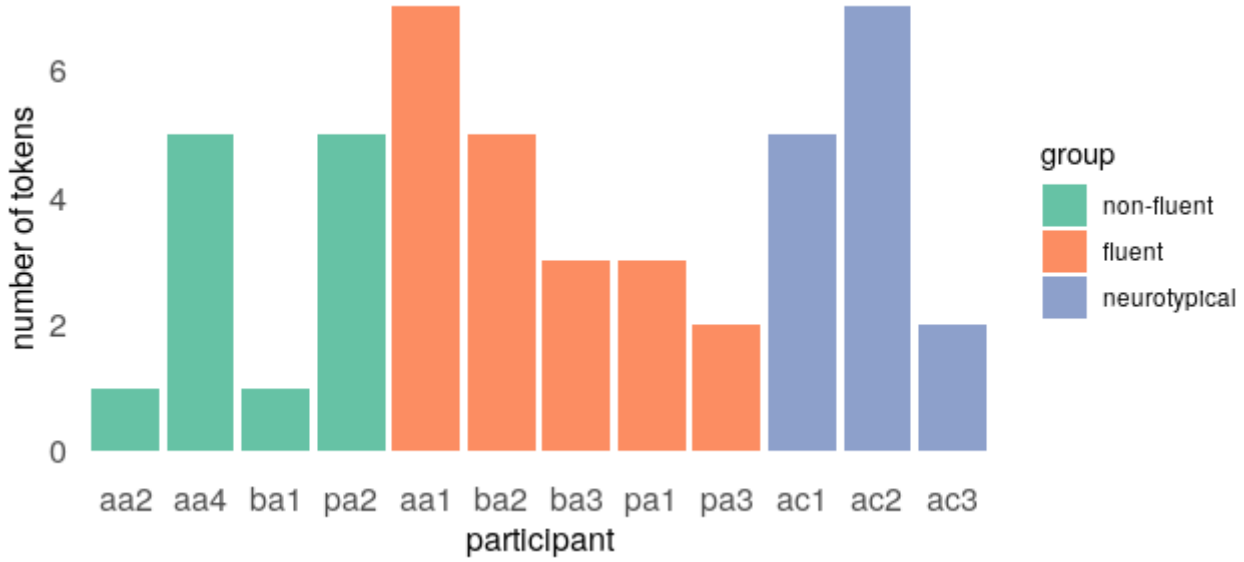

Figure 37: Number of inflected tokens with relative word form frequency lower than 0.05.

This pattern is consistent with aa3's language profile characterized by numerous anomias, an overall use of high frequency word and an overuse of pronouns, discourse management particles, and pronominal and adverbial demonstratives. It should be notice that participants pa3 and ba3 who have a similar general profile are also characterized by an overall higher relative frequencies such that pa3 has a median relative oral frequency of number-case of 0.223 and ba3 has a median relative syn2020 word form frequency of 0.275. Similarly, the production of ba1 is characterized by a low number of low frequency tokens (a single use of a word form with relative frequency below 5 %) and a relatively higher relative frequency overall (median oral number-case frequency of 0.227).[51]

We find a very different profile for participant aa4 who has nonfluent, relatively more severe aphasia (TMA). The median values of relative frequency on all three presented measures is 0.15 (0.158 for oral number-case frequency). When we look at the individual tokens, we find that these numbers are driven by five low frequency tokens, four of which are instances of the dative case which is one of the case with the lowest overall frequency in the corpora. The dative almost exclusively functions to express the recipient role in ditransitive verbs or as the case governed by the allative preposition *k* 'to(ward)'. This points to two things: one, this participant has access to the Czech transfer argument structure construction as well as the "allative" construction k N-DAT, which, when successfully retrieved, have a strong enough signal to facilitate access to these otherwise infrequent word forms, which are, however, strongly associated with these two constructions. A similar pattern can be observed for another nonfluent participant, pa2, who also produced a relatively high number of low frequency inflected tokens with the median values ranging from 0.101 for the oral number-case frequencies to 0.155 for SYN2020 word form frequencies. An inspection of the produced tokens reveals that three of the low frequency tokens are instances of the instrumental case. These all appear in adjacent c-units and all are governed by the preposition *před* combined with the verbs *prchat* 'flee', *schovat se* 'hide', and *utéct* 'run away'. The multifunctional preposition has an ablative

51 Median is used as the preferred measure of central tendency for the corpus data because it is more resistant to extreme values.

‘from’ meaning in these contexts. Given that all these instances can be taken to pertain to a single frame of escaping and the fact that they appear in close vicinity of each other, the successful production of these word forms can again be explained by a combined effect of a specific argument structure construction and priming.

The last pattern I want to discuss is aa2. Participant aa2 has language profile similar to aa4, recall that both participants cluster together based on the discourse production measures used in section 4.2.4. Their speech is characterized by short c-units with a minimum of verbs, frequent anomias and a high rate of disfluencies. A look at the summary in Table 24 shows a low number of inflected tokens in total, although the absolute number is higher than both aa3 and aa4 as well, being similar to pa1 and pa3. We also see relative frequencies that are comparable or slightly above the averages of the whole sample and the neurotypical participants. The median values for aa2, the neurotypical speakers, and all the participants combined is summarized in Table 25.

measure	aa2	neurotypical participants	all participants
number-case rel fq (oral)	0.186	0.152	0.158
number-case rel fq (syn2020)	0.158	0.15	0.144
word form rel fq (syn2020)	0.259	0.229	0.225

Table 25: Median values of selected frequency measures of inflected nouns for participant aa2, the neurotypical group, and all participants; rel fq = relative frequency.

A closer inspection of the individual tokens reveals an interesting pattern. As already discussed, aa2 had a high number of anomic pauses and errors in noun retrieval and these were oftentimes resolved in cooperative repairs with the administrator during the recording session. These repairs proceeded such that the administrator suggested a possible target which was either accepted or rejected by the participant. Accepted prompts resulted in the repetition of the suggested word. Figure 38 shows the relative frequencies for oral number-case and syn2020 word form. Interestingly, a number of the low frequency tokens are the ones that were prompted by the administrator rather than spontaneously retrieved.

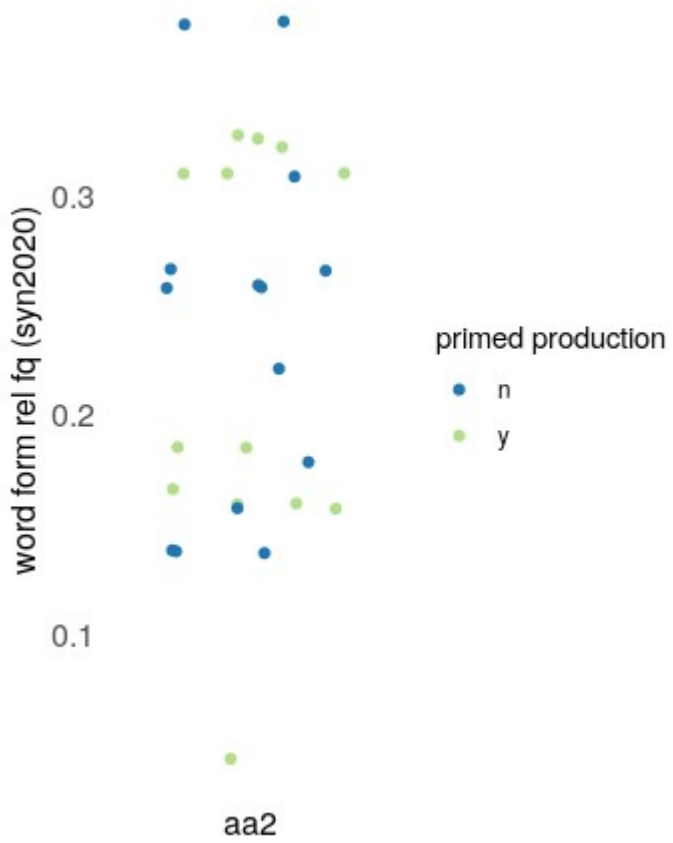

Figure 38: Relative word form frequencies of primed and unprimend tokens produced by participant aa2.

There is one additional factor that may play a role in this pattern and would also be relevant for the other participants. Three additional low frequency tokens produced by aa2 are plural nominative forms. There is a possibility that the pluralization construction [[R]-ı] facilitates the retrieval of such forms. This construction has the highest type and token frequency of all pluralization patterns. [[R]-ı] is the default for masculine and feminine nouns, while neuters do not appear in this construction, the category default being [[R]-a].[52] Figure 39 shows the distribution of plural nominative tokens in relation to oral number-case frequency.[53] When both of these factors are taken into account, we see that the majority of tokens spontaneously produced by aa2 are markedly above average (eight tokens with a median value of 0.264 for oral number-case and 0.216 for syn2020 number-case respectively).

52 Note that a structuralist analysis might arrive at a similar conclusion based on the fact that number is an inherent inflectional category, while case is a relational category governed prototypically by verbs and prepositions, but also by numerals and nouns in Czech.

53 Word form frequency data is not available in this case since some of the inflectional classes in Czech have form syncretic between singular and plural.

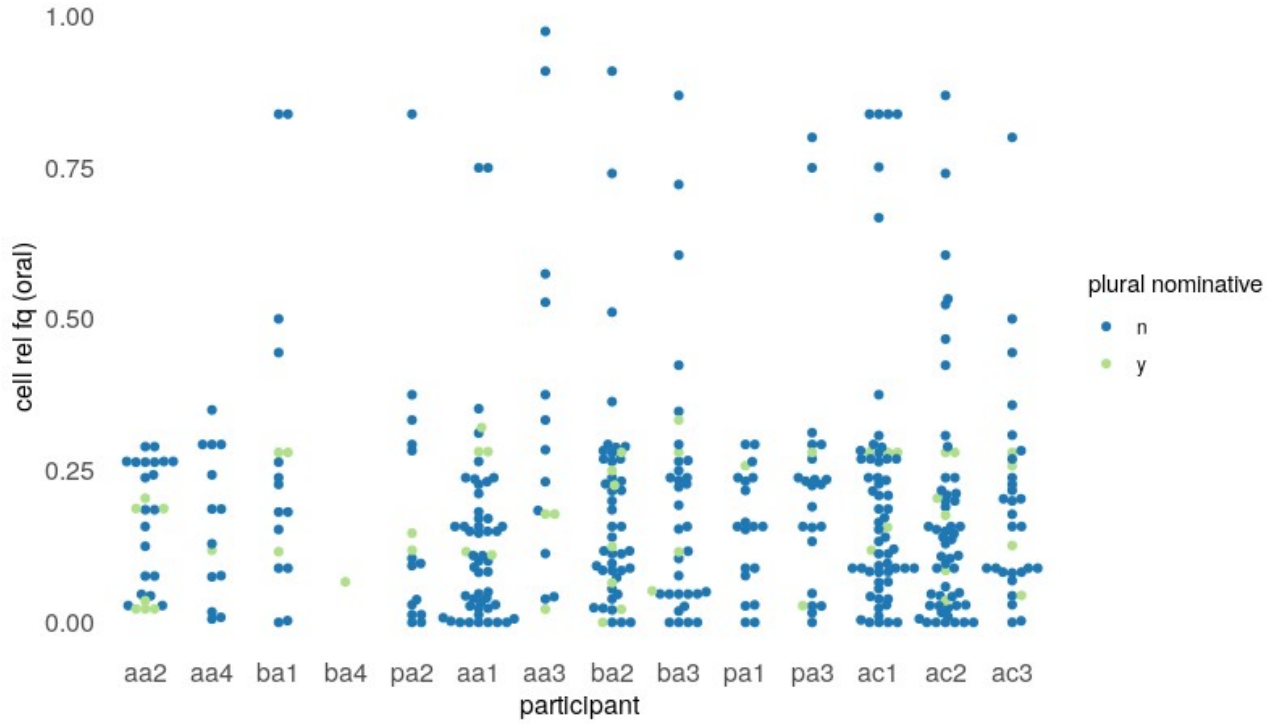

Figure 39: Relative frequencies of case-number combinations in oral of inflected tokens. Plural nominatives are marked in green.

Two important points can be concluded based on the analysis presented in this section. First, we saw that the non-fluent participants aa2 and aa4 produced a markedly lower number of inflected nouns, a similar pattern also occurred in aa3. This can be explained by a qualitatively different production strategy in these participants who rely on the existential-presentative construction. When relative ferqeuncy of number-case combinations and word forms is accounted for, we saw that all participants produce a high number of items with higher relative and rank frequency. However, there is also a trend in the expected direction such that participants with lower level of fluency and lower noun retrieval capacities produce comparatively lower numbers of tokens with low relative frequency and the case may be made that they rely more on high frequency items. Simultaneously, when such participants do produce rare forms it is in the context of particular constructions, such as the argument structure construction of escaping with the preposition *před* 'from, lit. in front of' and the tokens are found in close vicinity of each other, suggesting also the role of priming. Moreover, both low frequency inflected types produced by aa4 and pa2 have arguably salient forms ending in -ovi, used for masculine animate referents, in the dative and -(e)m in the instrumental, which is used for all masculine and neuter nouns. The ease of retrieval of the dative masculine animate forms may be further facilitated by the transfer argument structure construction in which recipients, prototypically sentient animate beings are marked by the dative. All of these findings again demonstrate the importance of the assumptions of the usage-based framework about language representation and processing. I hope to have provided in this section an example of the way in which a usage-based analysis can approach aphasic data and point to novel predictions and potential explanations with possible clinical implications.

4.4 Disfluencies as potential markers of chunkhood

Disfluencies, particularly silent pauses and hesitation sounds constitute an integral part of neurotypical spoken language. A body of mainly psycholiguistic research (but cf. Gilquin 2013 as an example of corpus linguistic approaches to this topic) focused mainly on silent pauses starting from the 1950s has repeatedly demonstrated that disfluencies are a marker of language planning. Researchers have focused on the distribution of disfluencies mainly in relation to clause structure but early on there has also been interest in the role of frequency and probability of (co)occurrence. For instance, Goldman Eisler has demonstrated in a series of pioneering studies that the predictability of words measured as transitional probability is correlated with pause occurrence and length (for representative examples see, e.g., 1958; 1961; or 1964).[54] Similarly, Lounsbury (1954) has hypothesized that the occurrence of pauses is strongly correlated to predictability and uncertainty. To provide a recent example, Kapatsinski (2010) has focused on the frequency characteristics of words in the context of replacement repairs, i.e. situations when speakers replace a word that has just been produced or whose production has commenced with a different word. He found that low frequency words are more prone to such repairs which he interprets in the usage-based framework as evidence for the claim that words with higher frequency are produced more automatically. Fraundorf and Watson (2014) analyzed the narratives of 15 English speakers and found that the occurrence of silent pauses was predicted by word frequency, while filled pauses (hesitations) were connected to the planning of whole utterances and occurred frequently before the introduction of a new plot point within the story. In a study that bears the most relevance to the present work, Schneider (2016) analyzed the distribution of silent pauses and hesitations in a corpus of spoken American English within the usage-based framework. Schneider focused on prepositional phrases as well as the elements preceding them and tested the hypothesis that the occurrence of disfluencies is indicative of chunking. Chunks are strings of words that can, but necessarily need not overlap with constituents and constructions and that are used frequently together which results in their being directly represented in language memory making them more accessible in processing (Zeschel 2008; Kapatsinski & Radicke 2009; Bybee 2010: chap. 3). Schneider used absolute frequency of cooccurrence of prepositions and noun complements as well as several association measures (forward and backward transitional probability, mutual information, and lexical gravity, a measure of how likely a specific combinations of words is among all possible combinations of words in a given bigram). She found that combinations that are more strongly associated have indeed a lesser probability of disfluency occurrence which can be interpreted as evidence of "chunkhood".

The role of pauses in particular and disfluencies in general has naturally been also studied in aphasia which is characterized by an increased frequency of disfluencies. However, the bulk of research has focused on the differences in pausing patterns between neurotypical speakers and speakers with aphasia.For instance, Feyereisen and colleagues (1986) showed that silence duration can be used as a measure of fluency in aphasia that is correlated to other measures in use. Similarly, Ellis and Rittman (2009) show that persons with aphasia following subcortical stroke produce more disfluencies than neurotypical speakers even several years post onset. In an analysis of connected speech of

54 Transitional probability is essentially the probability that when a speaker hears or sees the word A it will be followed by a given word B. Backward transitional probability which is also used in the present work expresses the probability that a given word B is preceded by the word A. This measure has been demonstrated to play a role in language acquisition and processing as an integral part of statistical learning and language prediction (e.g. Thompson & Newport 2007; Pelucchi, Hay & Saffran 2009; Onnis & Huettig 2021).

people with primary progressive aphasia (PPA), a neurodegenerative disease that selectively affects language, Mack and collaborators found that both neurologically healthy adults as well as persons with different PPA subtypes have a higher pause rate before low frequency words. Recently, Angelopoulou and colleagues have provided a detailed analysis of qualitative, quantitative, and distributional properties of silent pauses in Greek speakers with aphasia and neurotypical participants and found quantitative differences between the two groups. They describe two pause types, short and long pauses, that appear in both groups and have comparable distributions but differ in thresholds.[55] The aim of this section is to apply the usage-based framework on the disfluencies that accompany PPs in the corpus and to explore the role of frequency in this context.

4.4.1 Overall distribution of disfluencies in the corpus

Before I present a detailed description of disfluencies in the context of prepositional phrases with regards to possible effects of frequency and probabilistic processing, I describe a very general general picture of disfluent production in the subcorpus. The data presented in this section includes only the production of speakers with aphasia with the exclusion of participant ba4 based on the fact that this speaker produced a great number of single word utterances which logically implies that the majority of disfluencies produced by ba4 is located at c-unit boundaries.

4.4.1.1 Data preparation

The transcripts from the subcorpus were processed to string adjacent disfluencies (silent pauses, hesitations, and word fragments) into "disfluency chunks" such that, e.g., the sequence *<tut> a 0.5 <hes> s- 0.2 za nim se zavřely dveře* (aa1, c-unit 26) '<tut> and 0.5 c- 0.2 the door closed behind him' was transformed into *dfl a dfl za nim se zavřely dveře*.[56] Transformed in this way, the data was used to obtain a general picture of the distributional pattern of disfluencies within individual c-units. This analysis was only performed on the production of participants with aphasia. Participant ba4 was excluded from this analysis based on the fact that a substantial proportion of his c-units are single word utterances and thus not suitable for the analysis presented in this section.

Each disfluency chunk was annotated for preceding and following word including their grammatical characteristics. Two additional variables were added to this:

- place of disfluency: The locus was annotated as c-unit boundary when it occurred at the beginning or end of c-unit or a clause boundary when it occurred between a main clause and a subordinate clause or between two coordinated subordinate clauses. Disfluency chunk was classified as phrase boundary when it occurred between two phrases that were not in the relationship of a head and a dependent. A phrase internal chunk was coded when the disfluency occurred between a head and its immediate dependent. This coding was only used for

55 There is also a group of studies that have focused on pausing patterns in aphasia employing the Conversation analytical framework (Lesser & Milroy 1993; Smolash 1997; Sophie, Croteau & Tremblay-Beausejour 2007). This research has addressed the role of pauses in turn taking in conversations between speakers with aphasia and their significant others and found that particularly long pauses may result in turn loss in this context in contrast to conversations between neurotypical speakers.

56 The information regarding the different types of disfluencies in the chunk was retained, as will become apparent below.

directly adjacent phrases which effectively ruled out VPs. Verb-complex-internal disfluencies were annotated instead, marking such cases where the disfluency occurred between an auxiliary or a modal verb and a non-finite lexical verb or between a reflexive pronoun and the verb.

- success of target retrieval: Fully-formed word tokens were classified according to the success of retrieval such that abandoned clauses/c-units and repairs and reformulations were marked. Tokens without any clear indications of retracing were treated as successfully retrieved targets.

4.4.1.2 Data description

The aim of this section is to describe the general distributional characteristics of the disfluency chunks. I briefly summarize the position of disfluency chunks in c-units and the number of successful continuations following disfluencies. I then proceed with a detailed analysis of the relationship between disfluency occurrence and usage/performance factors in the context of prepositional phrases.

Table 26 provides an overview of the number of disfluency chunks for each participant, total number of c-units as well as the mean number of well-formed words in a c-unit and the mean number of disfluencies and fluent trigrams are repeated here from Section 4.2.1 for reference.

participant	fluency	number of disfluency chunks	number of c-units	number of tokens	MLU	mean number of fluent trigrams	disfluency chunks per word
aa1	fluent	135	103	790	7.515	3.427	0.178
aa2	nfluent	282	159	508	2.830	0.208	0.702
aa3	fluent	302	168	957	5.381	1.345	0.352
aa4	nfluent	105	77	225	2.714	0.286	0.560
ba1	nfluent	89	41	157	3.780	0.463	0.594
ba2	fluent	292	122	663	5.254	1.189	0.485
ba3	fluent	224	114	759	5.991	1.746	0.346
pa1	fluent	110	87	488	5.414	1.713	0.236
pa2	nfluent	51	51	221	4.235	1.196	0.241
pa3	fluent	192	106	806	7.057	2.311	0.270

Table 26: Summary of the total number of disfluent chunks produced by participants with regards to the measures of fluency.

Note that the fluent participants produce a considerable number of disfluency chunks. These are mostly connected to lexical retrieval problems and include silent and filled anomic pauses as well as retracing contexts in which the ongoing repair increases processing load. As already mentioned in #section, rather than the overall mean number of disfluency chunks per c-unit, it is the mean number of disfluency chunks per word that sets apart the nonfluent speakers where in the nonfluent group a disfluency on average occurs every two words as is also evident in the mean number of fluent trigrams per c-unit. As already discussed an exception in this regard is participant pa2 who is grouped together with the other nonfluent participants mainly based on reduced c-unit length.

When we look at the place of occurrence of disfluencies, this was categorized as c-unit boundary, subordinate clause boundary, phrase boundary, or phrase internal. The remaining disfluency chunks appear mostly in the context of unsuccessful retrieval and retracing. c-unit boundary position also includes cases where a disfluency chunk occurs between a c-unit-initial connective or discourse operator and the following clause as such/per se. Similarly, disfluencies located between a clause final word and a following discourse marker which is the final word of the c-unit were also coded as c-unit boundary positions. These cases are illustrated in 25.a and b respectively. Disfluencies that occur between a head and its immediate dependent as well as disfluencies in the "verb complex" were coded as phrase-internal (25.c). Cases where a disfluency occurs before or after a subordinating connective are marked as located on the subordinate clause boundary (25.d) and, finally, disfluencies that occur between two independent phrases are coded as phrase boundary in the sample, as shown in 25.e.

25. Examples of various disfluency positions within c-units

    a. *H 0.5 ale 0.5 ten lev 0.5 se zbudil 1* 'H 0.5 but the lion 0.5 woke up 1' (ba2: 81); note that a disfluency appears at both boundaries of the c-unit

    b. *tygr asi 0.2 jo 0.2 jo* 'a tiger probably 0.2 yeah 0.2 yeah' (pa1: 17)

    c. *tajdle sou 1 (p-) štyrý z- 3 pánové 0.2* 'there are 1 (m-) four z- 3 men 0.2' (pa2: 149); the token of interest here is the NP with a disfluency between the quantifier *čtyry* 'four' and the head noun *muži* 'men'

    d. *tady 0.2 kluk s holčičkou koukaji 0.2 co se teda bude dít 0.5* 'here 0.2 a boy with a girl are watching 0.2 what is going to happen 0.5' (pa1: 128); a disfluency occurs between the verb *koukat* 'watch' and the complement clause introduced with the connective *co* 'what, that'

    e. *a 0.5 H protože ten lev 0.2 byl takový 0.5 plachý ...* 'and 0.5 H because the lion 0.2 was sort of 0.5 shy ...' (ba2: 91); here the disfluency occurs between the adjectives forming a single adjective phrase *takový* 'such, sort of' and *plachý* 'shy'

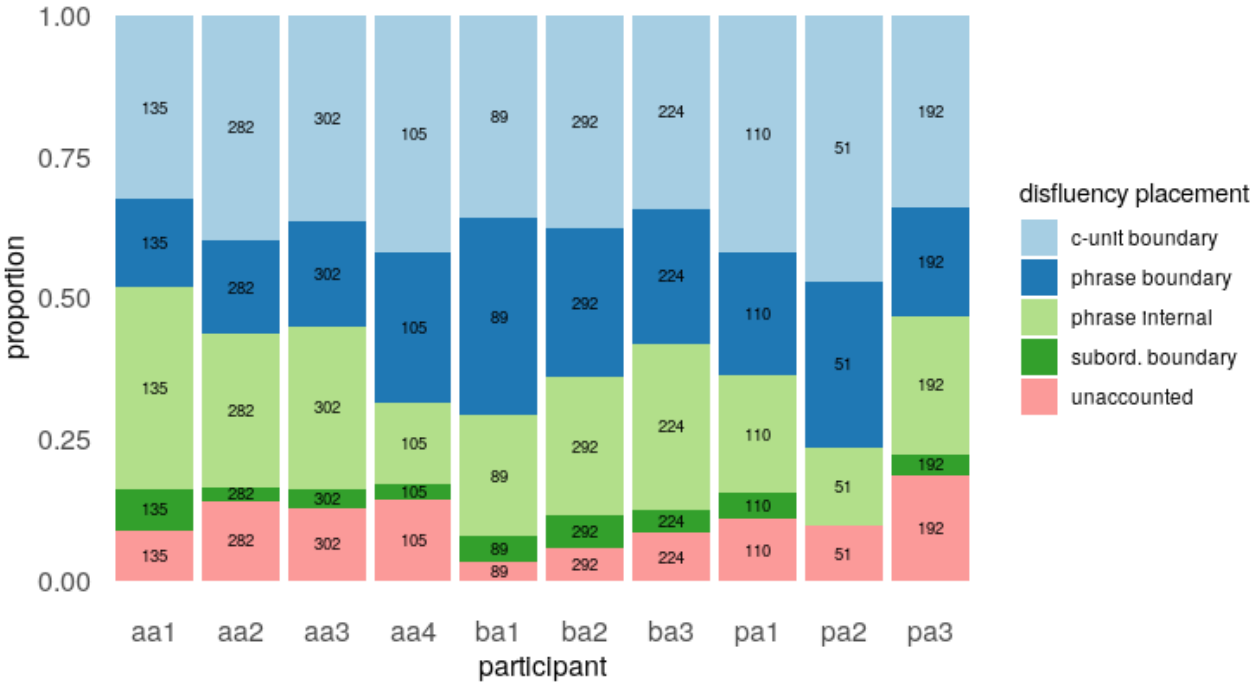

Figure 40: Proportion of disfluency placement with respect to clause structure; numbers in columns indicate counts.

Figure 40 shows the number and proportion of disfluency loci. We see that, across participants, the majority of disfluency chunks is spread more or less evenly between c-unit boundary, phrase boundary, and phrase internal position, while the proportion of clause boundary positions is very low. This follows from the fact there a relatively low number of subordinate clauses appears in the sample. c-unit boundary disfluencies are simultaneously more frequent than the phrase level disfluencies. Both c-unit-initial and -final position may reflect sentence planning due to the transcription scheme adopted in the design of the corpus. Where a (structurally and semantically) complete c-unit was followed by a silent pause which in turn was followed by a continuation of the participant's turn, such a pause was segmented as part of the former of the two consecutive c-units. c-unit-final disfluencies also include abandoned sentences, left unfinished by the participant, typically due to failed lexical retrieval. c-units which present a continuation within the turn and where a connective or a discourse marker appears at the onset and is followed by a disfluency chunk, may be interpreted as an interactional strategy of sort, albeit relatively infrequent, where the participant follows up on a c-unit with an immediate connective (most frequently the conjunction *a* 'and' or the discourse marker *tak* 'so, well') in order to keep the floor and only then plans the subsequent c-unit (26).

26. *ano smeták a 0.5 a 1 / tam 1 byl H ano H smeták* 'yes a broom and 0.5 and 1 / there 1 was H yes H a broom' (aa2: 257-258)

Two slightly different profiles seem to emerge in the data. On the one hand, we see a relatively lower proportion of phrase internal disfluencies in participants aa4 and pa2 with a proportion of 0.143 and 0.137 respectively, while sample mean of the remaining participants is 0.265 (sample mean = 0.24). This may have to do with the fact that both these participants produced relatively short c-units with below average number of backgrounded, modifying elements (mean number of modifiers for the group included in this analysis is 0.602, while aa4 produced on average 0.377 and

pa2 0.49 modifiers). This suggests that the syntactic complexity of these two participants' c-units is lower and there are fewer relevant places in which such a disfluency can occur. This is complemented by the proportion of phrase boundary disfluencies which is also slightly above average in both participants (aa4 = 0.267, pa2 = 0..294, mean of the remaining participants = 0.221, sample mean = 0.233). A similar profile would be expected in participant aa2 based on his similar overall language profile, whereas there is a considerable number of phrase internal disfluencies. However, on closer inspection, these are in the majority retracings of nouns and verbs marking lexical retrieval difficulties and a different strategy of word finding. We find a mirror image of sorts for participants aa1, aa3, and, to an extent, ba3 and pa3 who share a general language profile of longer c-units with an average to above average use of modifying elements, as summarized in Table 27. Furthermore, we see slightly higher proportions of disfluency occurrences in the "other" category in participants aa2, aa3, aa4, and pa3. This is an expected pattern given that aa2 and aa4 have a largely reduced level of fluency on the one hand and aa3 and pa3 share a language profile characterized by a reduced content load and a high number of lexical retrieval problems.

participant	mean modifiers	proportion phrase boundary (n)	proportion phrase internal (n)
aa1	0.932	0.156 (21)	0.356 (48)
aa3	0.589	0.185 (56)	0.288 (87)
ba3	0.763	0.237 (53)	0.295 (66)
pa3	1.02	0.193 (37)	0.245 (47)

Table 27: Proportion of pre-phrasal and phrase-internal disfluencies with respect to the number of modifiers produced by selected participants.

When we turn to the success of retrieval following a disfluency chunk, there are three broad categories that were already discussed above: successful retrieval, retracing, and no retrieval. It should be noticed that disfluencies that occurred on c-unit or clause boundary were not coded for success of retrieval with the exception of c-unit-final disfluency chunks in abandoned sentences. Figure 41 summarizes the counts and proportions of these three types for individual participants.

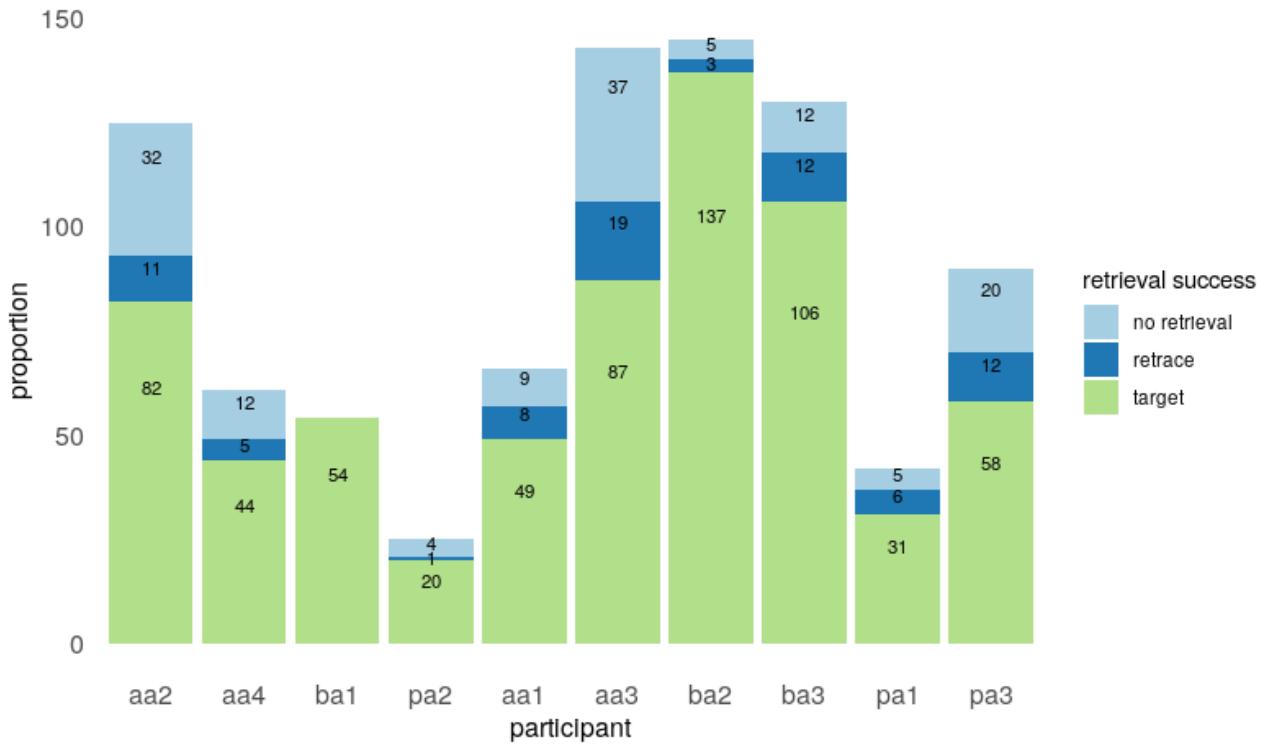

Figure 41: Proportion of disfluencies with respect to success of retrieval of the target word; numbers in columns indicate counts.

The observed patterns are expected based on the general language profiles of the individual participants. We see a relatively higher proportion of failed retrieval in participants aa2 and aa4 as well as a higher proportion of failed retrieval and retracing in aa3 and aa4. The production of aa2 and aa4 is characterized by a high number of sentence fragments and lexical retrieval errors/problems predicts this pattern where word finding difficulties frequently result in abandoned sentences, i.e. sentence fragments. Interestingly, aa3 and pa3 share this profile of frequent retrieval errors and sentence fragments with aa2 and aa4. However, their comparatively higher level of fluency results in more frequent retracing which brings them closer in this respect to aa1 and pa1. As noted, these two participants with anomic aphasia are subjectively the mildest cases in the sample and it is thus not surprising that when word finding difficulties were encountered by these participants they were more successful in repairs than the other participants.

4.4.2 A closer look: Prepositional phrases

The group of prepositional phrases was selected to examine the distribution of disfluencies in more detail, employing the usage-based perspective. This decision was in part motivated by Schneider (2016)'s study of PPs in spoken American English discussed above. Furthermore, there were two other reasons for selecting this group of constructions. Firstly, the majority of Czech prepositions governs cases that have word forms different from citation form and, as already discussed in section 4.3.2, this may be expected to require extra processing resources. Secondly, PPs were chosen for the analysis for practical reasons, since they are easily delimited and extracted from the corpus. Thirdly, PPs in Czech are oftentimes used in non-core semantic roles and as adjuncts which again can be loosely connected to increased syntactic complexity and, consequently, increased processing load.

Disfluencies may thus be expected to occur in the production of these forms. This is also what we see when we look at the occurrence of disfluency chunks in phrase boundary and phrase internal position in relation to part of speech and phrase type. Figure 42 which summarizes these distributions shows that all participants produced a number of disfluencies both before a preposition and within prepositional phrases. The aim of the analysis was to explore the role of frequency and syntactic probability in the occurrence of disfluencies in the context of these constructions.

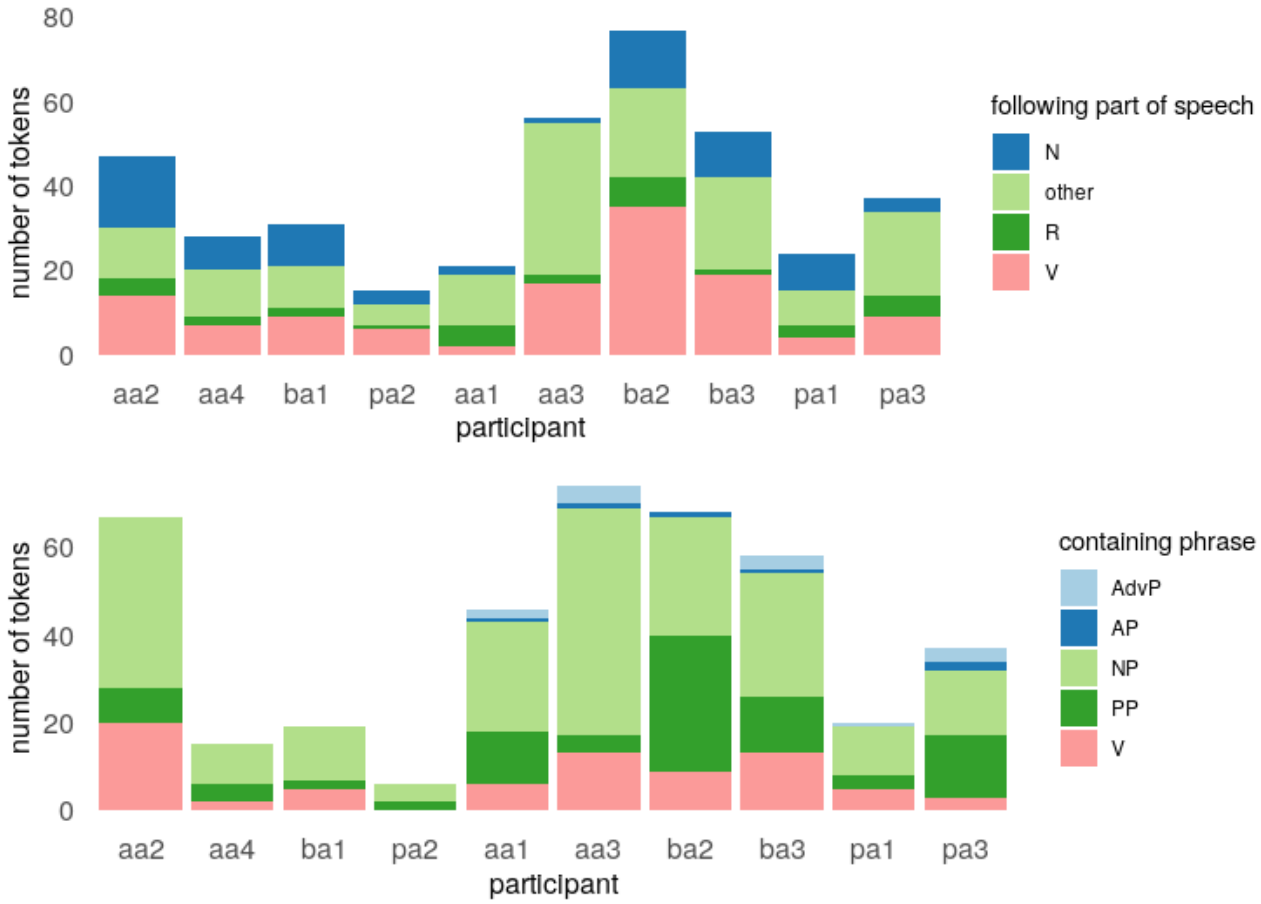

Figure 42: Number of disfluencies produced pre-phrasally (upper panel) and phrase-internally (lower panel) with respect to phrase structure.

The data from the subcorpus included a total of 278 PPs out of which 197 had a nominal complement, including three proper noun tokens that were excluded from the analysis. Similar to the analysis of nominal inflections in the previous section, all paraphasic word forms with clearly reconstructable targets were kept for the analysis as long as the paraphasic token shared the inflectional marker/paradigm with the target. The PP *v tomectom <frag> 0.2 štoudlíku* 'in this <frag> 0.2 vat' produced by ba3 was the only token excluded for these reasons. The distribution of complement part of speech is summarized in Figure 43. Only 15 cases include unsuccessful production of a PP with no complement retrieved, which is spread across seven of the ten participants (ba4 was excluded from the analysis as he did not produce any PPs in the subcorpus tasks). Four tokens had a non-nominal and non-pronominal complement which comprised two PPs with an adverbial complement: aa2's *na tam* 'on there', which is ungrammatical in Czech, and pa1's *za dřív* 'during earlier times, lit. during once, earlier', which is borderline acceptable at least. It is thus apparent that while

the PPs in the subcorpus did induce disfluent production, participants were overall relatively successful in their production. To allow for a more comprehensive/reliable analysis, the subcorpus data was complemented with data from the other part of the corpus which contributed additional 511 PP tokens such that the resulting sample included a total of 704 PPs with generic nominal complements; this sample consists of 510 unique combinations of prepositions and complement word forms.

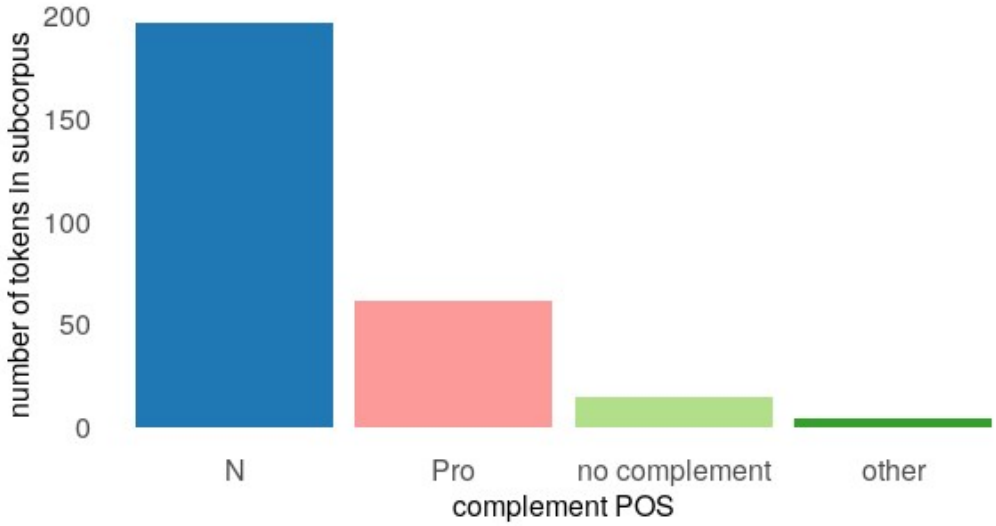

Figure 43: Number of PP complements in the corpus.

These PPs were annotated for a suite of variables that were expected to have a potential role in/contribute to disfluency occurrence.

- complement lemma and word form
- complement number and case
- preposition lemma
- whether the word form is different from the citation form
- fluency: Silent pauses, hesitation sounds, word fragments, and repetitions and reformulations were counted as disfluencies
- place of occurrence of disfluency: Disfluencies were coded as pre-phrasal if the disfluency preceded the preposition, a phrase-internal disfluency was used for occurrences between the preposition and the complement, a disfluency was classified as mixed when it occurred in both places. In cases where a modifier was present, position relative to the modifier was further noted.
- presence of any prenominal modifier: Prenominal modifiers were classified according to function into following categories: adjective, demonstrative, possessive, indefinite, and in-

terrogative. Where multiple modifier were used, all categories were marked. PPs with post-nominal modifiers were also marked.

- whether the PP is an argumnet of the clause predicate: Where the head of the PP was a verb, its argument status was annotated according to the valency dictionary of Czech. The entries in Vallex (Lopatková et al. 2020) differentiate between argument on the one hand and typical and optional non-argument complements and modifiers. These categories were also used in the annotation with one additional category for copulas
- frequency of the preposition
- cumulative lemma frequency of the complement noun
- frequency of complement word form
- frequency of complement number and case combination
- relative frequency of word form and number-case combination
- frequency of the PP bigram
- relative frequency of the PP out of the total frequency of the paradigmatic cell
- forward and backward transitional probability of the seqeunce P N: To account for the probability of cooccurrence of the preposition and the complement, forward and backward transitional probability was computed based on the corpus frequencies of the prepositions, the nouns, and the P N bigrams.

The frequency data was obtained from three corpora of Czech, a corpus of written language, a spoken corpus, and a corpus of movie subtitles. An exception to this was the word form frequency in the spoken corpus. This count was excluded because of the phonetic and morphological variation in spoken Czech. Word forms were regularized to orthographic variants before querying the corpora.

4.4.2.1 Assessment of item difficulty/processing demand

While the individual frequency measures might be expected to play a role in the representation and retrieval of the PPs, it is also to be expected that these measures may interact, under the usage-based model, since “coalitions” of features the effects of which may be relatively weak in individually but have a strong effect on the way language is used when combined have been demonstrated to be prevalent in language. Furthermore, different corpora have been demonstrated to produce more or less different predictions. In order to account for this, I devised a method to assess the potential difficulty or processing demand of the PPs for language production.

Under the assumptions of the usage-based model, high lemma frequency, high relative frequency of a given word form, and the strength of association/frequency of cooccurrence of a given noun with a particular preposition all contribute to the strength of memory trace and ease of retrieval. Nouns that have a high cumulative frequency have been indeed repeatedly demonstrated to be retrieved more easily and faster. There is also ample evidence that relative frequency of inflected forms facili-

tates processing. Moreover, inflected forms that are used more often are predicted to be stored as ready made chunks rather than retrieved as a superimposition of two partially filled sub-word schemas. The same is true about a multiword expression such as a particular preposition combined with a noun. While prepositions as such are "promiscuous" in the sense that there are only very general combinatorial restrictions with regards to semantics, e.g. *k* 'toward' is prototypically used with spatial nouns with an allative meaning. On the other hand, some nouns are used with particular prepositions with a relatively high frequency.[57] This is also reflected by the transitional probabilities in the data in that the third quartile is 0.001 for forward transitional probabilities based on the subtitle corpus, the same value is 0.141 for backward transitional probability. It could be hypothesized that there would be no strong associations between a particular preposition and lexically specific noun complements in a cloze task, while it would be much easier for Czech speakers to "guess" a preposition given an inflected noun. It follows from these assumptions that target PPs that have a higher frequency/probability may be expected to be easier to produce and result in fewer disfluencies.

In order to identify complement nouns and PPs that would have an advantage in production I used hierarchical clustering to obtain a corpus driven grouping of the PP items. This was done rather than focusing on the contribution of the individual frequency measures. The frequency measures were first submitted to principal component analysis and a hierarchical cluster analysis was subsequently performed on principal components. This analysis identified four clusters of PP types. Table 28 presents basic characteristics of these clusters, the clustering solution is shown in Figure 44. We see a partial overlap of PP types assigned to clusters 1 and 3. Cluster 2 is a small, clearly delimited grouping of just six noun lemmas with a very high frequency such as *den* 'day' or *rok* 'year' that are combined with different prepositions. In cluster 3, we see a number of PP types that lie close to clusters 1 and 3. There is also a group of, in a sense, outlier items that consist of extremely frequent and strongly associated PPs with a lower degree of compositionality, such as *v podstatě* 'in fact' or *v sobotu* 'on Saturday'. Clusters 2 and, in particular, 4 are potentially the most interesting with regards to the assumptions of the usage-based model. As already mentioned, cluster 2 is defined mostly by complements with very high cumulative lemma frequencies, it can also be seen that these combinations include prepositions with comparatively lower frequency. While the frequency of the prepositions that appear in the PP types in cluster 2 are relatively lower, we see that the forward transitional probability of these P N combinations is slightly higher in this cluster with the mean and median being almost an order of magnitude higher compared to the other clusters. Cluster 4 contains P N combinations which are characterized by a high number-case relative frequency of the complement and very high backward transitional probabilities. It also follows from these two measures that a very high proportion of the total frequency of a given number-case paradigmatic cell is accounted for by the corresponding PP in this cluster. Clusters 1 and 3 contain the intermediary PP types with cluster 1 being characterized by a relatively lower frequency of prepositions and cluster 3 having higher cell relative frequencies as well as backward transitional probabilities.

57 For instance, the three most frequent complements of the preposition *k* account for only about 6 % of total occurrences of all PPs headed by *k*. Conversely, when one looks at the group of prepositions that govern the genitive case and their complements, there are nouns/word forms that occur almost exclusively with just a single preposition, e.g. *důchod* 'pension' occurs in the genitive form *důchodu* almost exclusively with the preposition *do* 'in, to' (93 %).

	cluster 1	cluster 2	cluster 3	cluster 4
P N count	251	15	185	57
N lemma fq (mean; median)	98; 24	1373; 1329	113; 24	122; 42
P lemma fq (mean; median)	4027; 4789	3058; 1544	13285; 14055	10843; 11259
cell rel fq (mean; median)	0.137; 0.103	0.176; 0.127	0.195; 0.188	0.449; 0.397
PP of cell (mean; median)	0.283; 0.176	0.306; 0.097	0.442; 0.405	0.813; 0.915
tpB (mean; median)	0.0326; 0.0149	0.0355; 0.0122	0.0789; 0.0533	0.3466; 0.3082
tpF (mean; median)	0.000807; 0.000103	0.022496; 0.007217	0.000482; 0.000079	0.004783; 0.001238

Table 28: Basic characteristics of the clusters of PP types; lemma frequency in pmw was rounded to whole numbers.

Figure 44: Clustering of individual PP tokens based on frequency measures.

The data based clustering method was complemented by an additional index. Using the frequency data from the subtitle corpus, which was used as a compromise between ORAL and SYN2020,[58] I obtained the standard measures of central tendency and dispersion for the measures that were also submitted to the cluster analysis and coded all values above the third quartile as high frequency. All values below the first quartile were similarly coded as low frequency. Category memberships across measures was then summed such that, for example, the PP *z konve* 'from a kettle/can' has one low frequency point for cumulative lemma frequency and three high frequency points for number-case relative frequency, proportion of PP within paradigmatic cell, and backward transitional probability.

We can now turn to the production of the PPs in the corpus and the distribution of disfluencies in this sample. Before I proceeded with the analysis, I removed additional 41 tokens for two reasons. One aa2's and two aa3's tokens were combinations of a preposition with a word form in an ungrammatical case, as shown in Table 29. Additional 38 tokens were removed because they were immediate direct repetitions of administrator prompts and their having been produced fluently is more likely explained by priming effects rather than successful spontaneous retrieval. The majority of such tokens was produced by aa2 (21), while the remainder was distributed evenly across all participants except for pa3.

participantt	PP produced	gloss	target PP	gloss
aa2: 355	s cukr	with sugar-sg.nom/acc	s cukrem	with sugar-sg.ins
aa3: 19	vo začatek	about beginning-sg.nom/acc	od začátku(?)	from beginning-sg.gen
aa3: 578	v češtinou	in Czech-sg.ins	s češtinou / v češtině	with Czech-sg.ins / in Czech-sg.loc

Table 29: Agrammatic PPs in the sample. Two target PPs for aa3: 578 present the appropriate complement form given the preposition that was produced and a preposition that governs the case produced with the noun respectively.

Table 30 summarizes the total number of PP tokens produced by participants as well as the number of fluently and disfluently produced tokens. These numbers correspond to the profiles discussed throughout the present work. Participants with reduced levels of fluency, aa2, aa4, ba1, and pa2 produced low numbers of PP tokens. aa2, aa4, and ba1 produced less than 50 % tokens fluently. The group of fluent patients aa3, ba3, and pa3 is again characterized by a lower number of total PPs produced with a relatively high proportion of disfluencies, consistent with their language profiles. The participants with anomic aphasia high numbers of PPs which are relatively fluent in the case of aa1 and pa1. However, there is also a number of disfluencies, consistent with a considerable number of word finding difficulties expected based on their diagnosis.

58 See also Brysbaert and New (2009) who show that frequency data based on a corpus of movie subtitles better predicts standard psycholinguistic measures of word recognition.

participant	fluent	disfluent	total	proportion fluent tokens
aa1	154	91	245	0.629
aa2	5	15	20	0.250
aa3	32	24	56	0.571
aa4	2	6	8	0.250
ba1	14	17	31	0.452
ba2	33	69	102	0.324
ba3	11	28	39	0.282
pa1	65	43	108	0.602
pa2	15	6	21	0.714
pa3	11	22	33	0.333

Table 30: Summary of PPs produced by participants fluently and with disfluencies.

As discussed in the description of data processing steps and annotation scheme, the main goal of this analysis was to explore the role of both structural and usage factors in the distribution of disfluencies. Structural complexity is represented by the number-case value of the complement as well as "inflectedness" of the word form, i.e. whether it differs formally from the citation form. In addition, the presence of any prenominal modifiers occurring between the preposition and the complement as well as the role of the PP in the argument structure of a given c-unit were also considered. Cluster membership and the high frequency scores represent usage/performance factors.

These factors were not only explored in relation to fluency as such but also the place of occurrence of disfluency pre-phrasally or phrase-internally. This was motivated by the assumption of the usage-based model with respect to chunking and probabilistic processing. In simplified terms, words that occur frequently together may form chunks or prefabs that may but need not overlap with constructs, similar to morphologically complex words the retrieval of which is modulated by frequency. If a PP is produced with a prephrasal, rather than phrase-internal disfluency, this might be taken as a sign of stronger association between the preposition and the complement and, consequently, their forming a chunk and being retrieved as a whole. Figures 45 and 46 show the number of pre-phrasal and phrase-internal disfluencies overall and for individual participants. The number of pre-phrasal disfluencies is relatively high. This pattern is particularly interesting when compared with the literature using neurotypical speech. For instance, Maclay and Osgood (1959) analyzed the distributional properties of different types of disfluencies in samples of spontanous speech of English native speakers and found that phrase-internal filled and silent pauses were approximately twice as frequent as pre-phrasal disfluencies for PPs. Similarly, in his typologically based study of the suffixa-

tion preference in the languages of the world, Himmelmann (2014) shows on spoken data from English and German that disfluencies tend to occur more frequently between a function word and a host, i.e., in the present work, a preposition and a complement noun. For instance, of has only four pre-phrasal disfluencies out of 26 and in five out of 30 (Himmelmann 2014: 938). The pattern found in the present analysis is, on the other hand, similar to Schneider (2016)'s study that investigated PPs specifically and found a comparable number of disfluencies in the pre-phrasal and phrase-internal position, however Schneider did not focus on possible differences in disfluency placement. The data will be presented as fluent and pre-phrasally or phrase-internally disfluent in the remainder of the section in order to explore potential differencies between these two types.

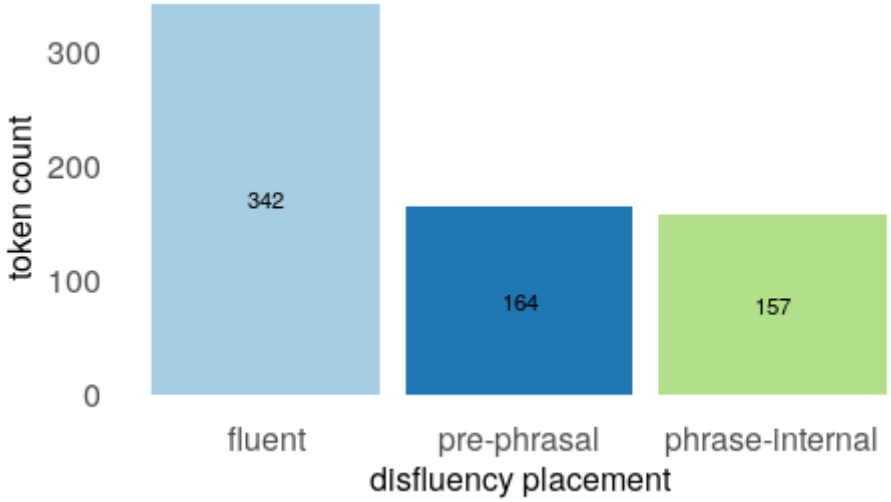

Figure 45: Number of PP tokens with respect to fluency.

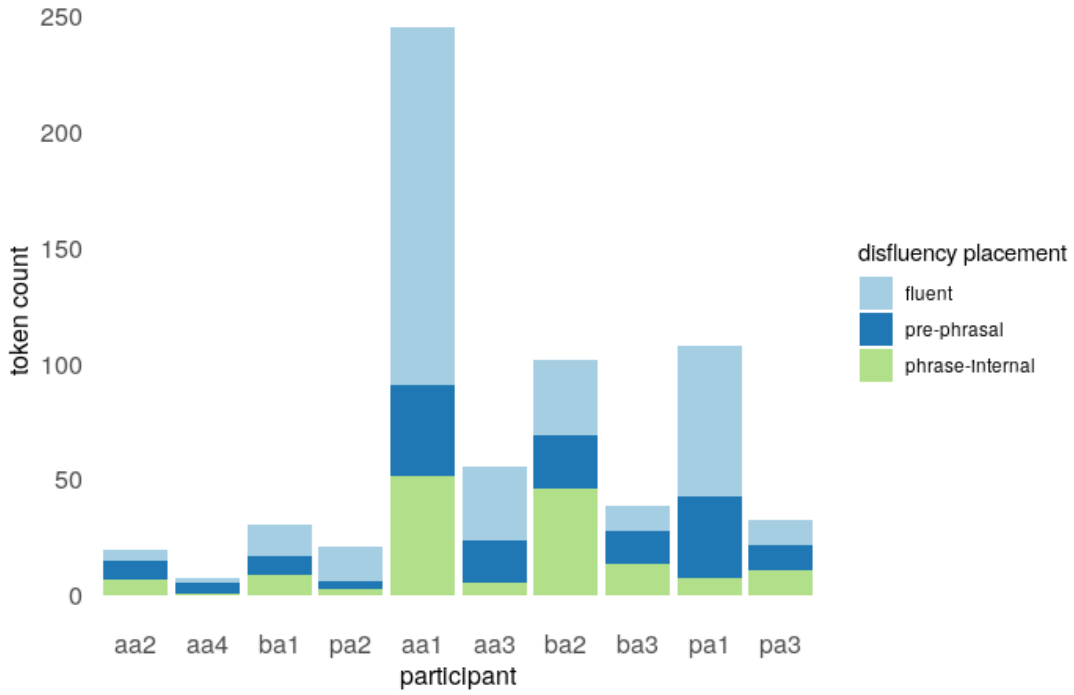

Figure 46: Number of PP tokens with respect to fluency produced by individual participants.

Figure 47shows the distribution of disfluencies with regards to argumenthood, "inflectedness", and case and number configuration of the complement for the whole sample. Particular number-case combinations were also complemented by an auxiliary variable contrasting the accusative and the genitive as more "syntactic" cases with the "semantic" cases.[59] It is clearly visible that the distribution of disfluencies is comparable for all these factors. This does not change when individual differences are included/taken into account. Slightly higher fluency rate visible for the plural dative and locative, for instance, is explained by the fact that the majority of such tokens was produced by the most fluent participant aa1 and pa1.

59 The genitive was grouped with the accusative as it serves to express possessors (*auto kamarádky* 'car of a friend') and also core arguments in nominalizations as in, e.g., *příchod kamarádky* 'arrival of a friend' or *prodej auta* 'sale of a car'

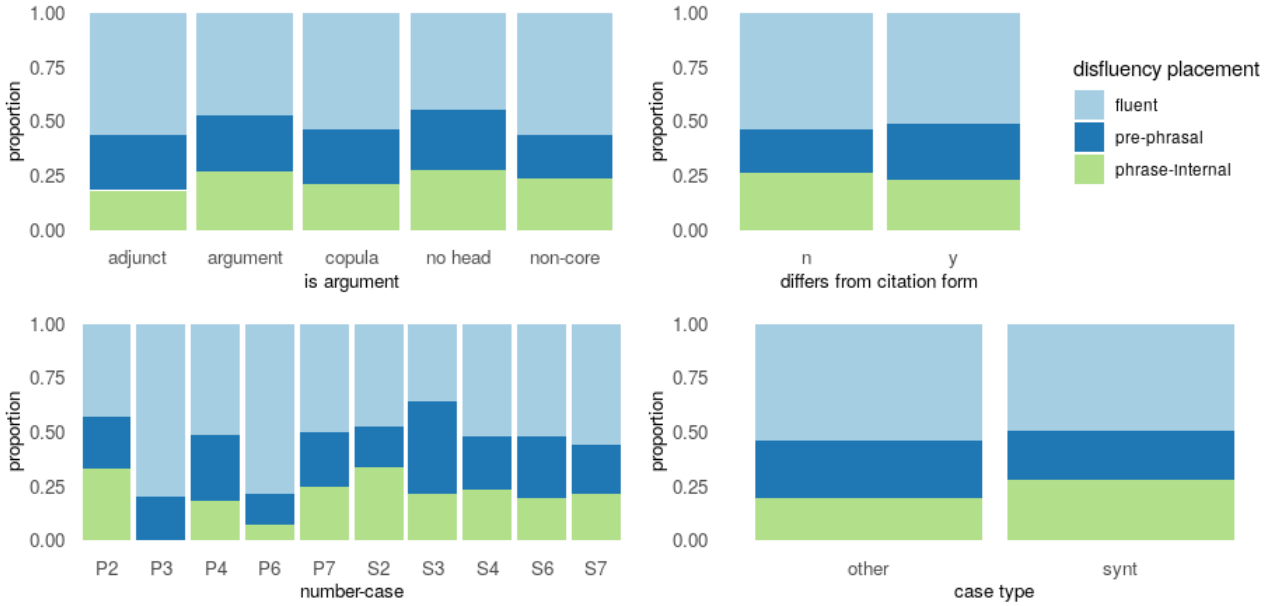

Figure 47: Proportions of disfluency placement in PP tokens with respect to argumenthood (upper left hand panel), inflectedness of complement (upper right hand panel), number-case values (lower left hand panel), and grammatical and semantic cases (lower left hand panel).

The remaining structural variable is modifier presence. Table 31 summarizes the number and type of modifiers occurring in the sample across all participants. Approximately 40 % of all tokens include a prenominal modifier with demonstrative pronouns accounting for the majority of all occurrences (ca. 63 %). This is not surprising given the fact that demonstratives have been argued to be in the process of grammaticalization in definite articles (Zíková 2018).

modifier type	frequency
none	404
demonstrative	164
adjective	43
multiple	23
indefinite	14
interrogative	11
possessive	4

Table 31: Distribution of prenominal modifiers in the sample.

Figure 48 show the distribution of disfluencies in relation to modifier presence. #figure shows the relationship between disfluencies and modifier presence and type; modifiers are non-demonstrative

modifiers are grouped together. The aggregated data shows that PPs with no modifiers have a higher fluency rate (57 %) as well as a lower proportion of phrase-internal disfluencies (13 % phrase-internally and 30 % pre-phrasally). The presence of a modifier makes the PP more complex and longer which creates more "slots" for potential disfluency. On the other hand, there is no difference between demonstratives and all the other modifiers, suggesting the more grammaticalized are in this case not less demanding during processing. One might, conversely, argue that adjectival modifiers might produce more disfluencies since they are more lexical and contentful and, moreover, an overwhelming majority of adjectives agrees in gender, number, and case. This pattern is, however, not supported by this particular sample as the proportion of phrase-internal disfluencies is similar (42 % for adjectival and 40 % for pronominal modifier respectively). #figure shows that this pattern holds when individual differences are taken into account. We see a higher number of fluently produced PPs in absence of modifiers and, conversely, a higher rate of phrase-internal disfluencies in modified PPs. A small exception to this pattern is participant aa2. However, he produced only two modified tokens and no conclusions can thus be drawn from this. Furthermore, one of these tokens is *na státní silnici* 'on the state road' which is a technical term referring to a specific type of roadway. The adjective *státní* has the same form in all genders and across the paradigmatic cells.

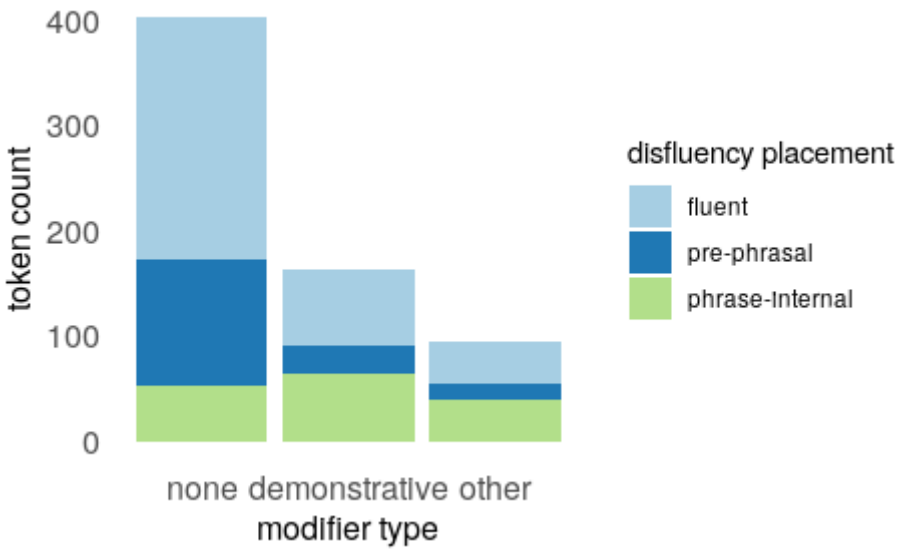

Figure 48: Number of unmodified tokens and tokens modified by a demonstrative or a different modifier with respect to fluency.

I turn now to the role of the frequency measures, splitting the data between modified and unmodified PPs. Figure 49 shows a general distribution of disfluencies for the modified and unmodified PPs in relation to their membership in the clusters described above. PPs belonging to clusters 2 and 4 that were predicted to facilitate production are grouped and compared to clusters 1 and 3.

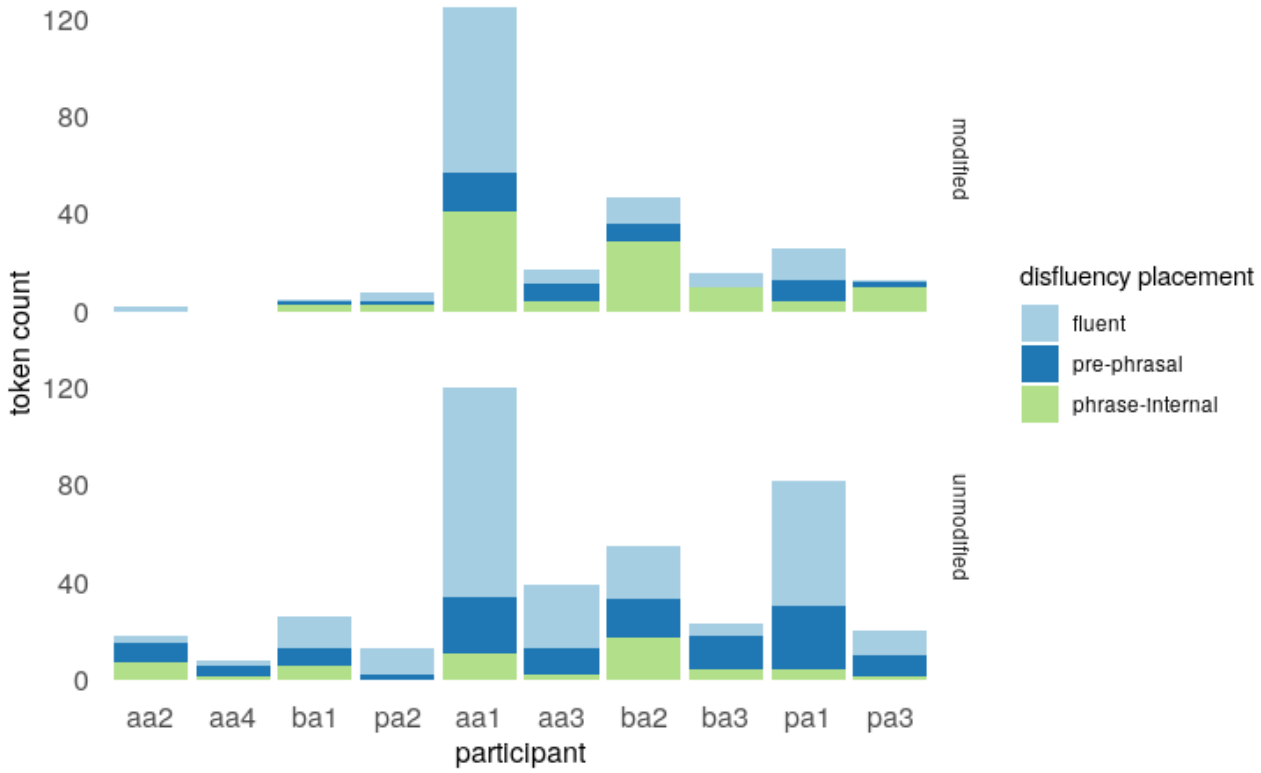

Figure 49: Disfluency placement in modified and unmodified PPs produced by individual participants.

While it was already established that modified PPs are on the whole less fluent and have a higher rate of phrase-internal disfluencies, this figure suggests an added effect of frequency of use. When we look at the fluency rates in the unmodified PPs (upper panel), the PPs in the high frequency clusters are generally more fluent; 59 of 87 (67.82 %) PPs in clusters 2 and 4 were produced fluently, while 171 of 317 (53.94 %) PPs in clusters 1 and 3 were fluent. Furthermore, if we focus on the disfluency location, only four of the 28 (14.29 %) disfluently produced PPs in clusters 2 and 4 occur phrase-internally, whereas 49 of the 146 (33.56 %) disfluent PPs in clusters 1 and 3 contain a phrase-internal disfluency. A similar pattern is found in modified PPs: There are 31 fluent PPs of the total 46 (67.39 %) belonging to clusters 2 and 4, while only 81 out of 213 (38.03 %) PP in clusters 1 and 3 were produced fluently. These patterns suggest that PPs that are used more frequently are on the whole easier to retrieve and induce fewer disfluencies in the corpus. However, it is also important to take potential individual differences into account, given that the participants have very different language profiles.

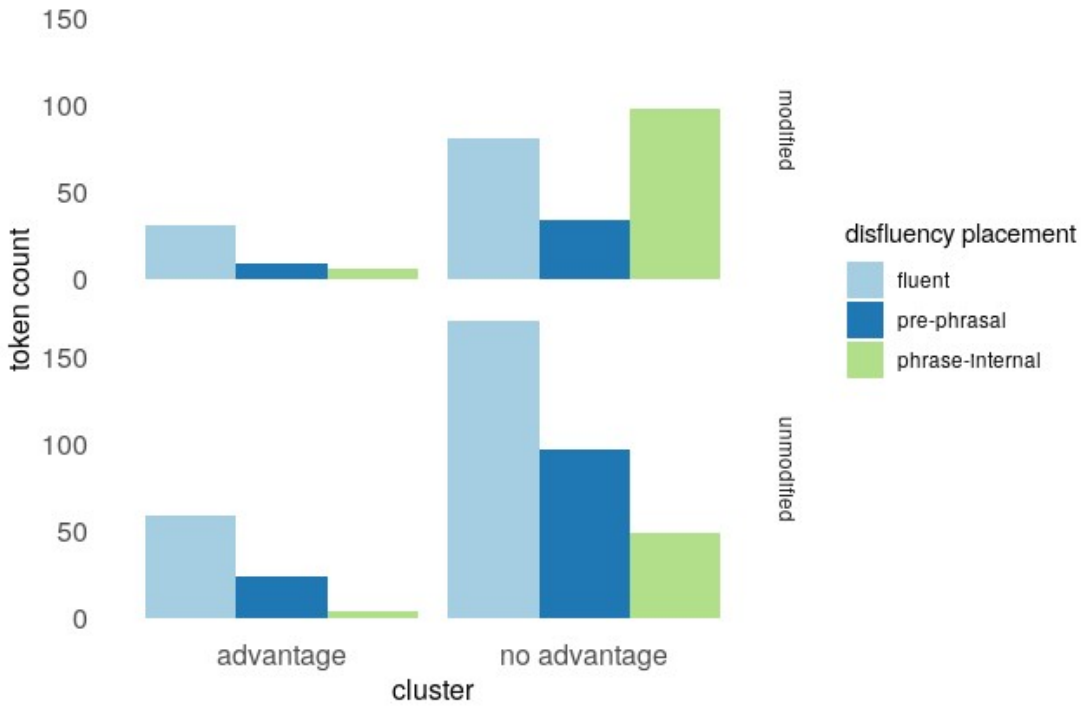

Figure 50: Disfluency placement in modified and unmodified PPs with respect to cluster membership.

Beginning with the unmodified data, the distribution of disfluencies overll is shown in Figure 50 while the numbers for individual participants is presented in Figure 51. The number of tokens belonging to the respective pairs of clusters produced by some of the participants makes comparisons at individual level very difficult. This is particularly the case for participants aa2, aa4, and pa3. However, a visual inspection of the data shows that where comparison is possible, the general trend is to an extent visible also at individual level. The disfluency rate is on the whole relatively higher for tokens in clusters 1 and 3, especially when one focuses on the proportion of pre-phrasal and phrase-initial occurrences, we see that phrase-initial disfluencies are also more frequent in clusters 1 and 3.

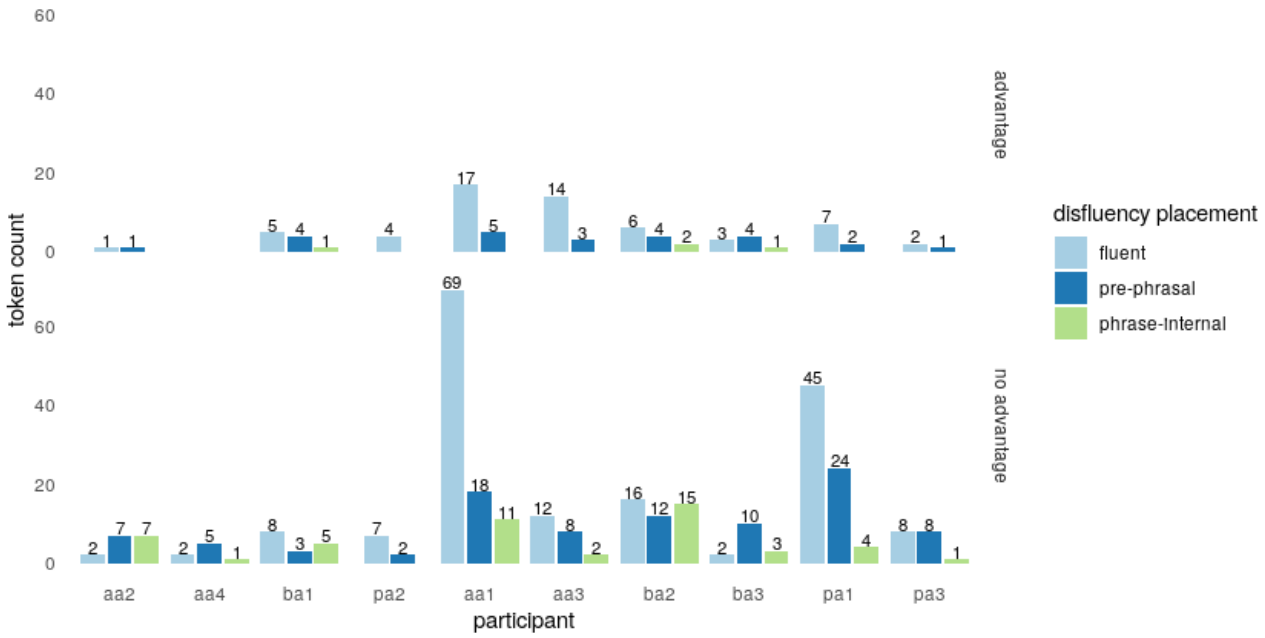

Figure 51: Disfluency placement of PP produced by participants with respect to cluster membership.

Table 32 summarizes the production of PPs with regard to cluster membership for participants aa1, aa3, ba1, ba2, ba3, and pa1. We see that fluency rates are generally higher for tokens from clusters 2 and 4. Furthermore, the proportion of phrase-internal disfluencies is higher for clusters 1 and 3 in all participants. However, it is important to stress again that some of the numbers are very low and these really are to be conceived of as suggesting a trend rather than a strong pattern. It is also worth mentioning that the single disfluent tokens produced by participants aa2 and pa3 belonging to cluster 4 were produced with a pre-phrasal disfluency and, similarly, the four cluster 4 tokens of pa2 were produced fluently. aa2 produced one fully fluent and one pre-phrasally disfluent token of *v kleci* 'in the cage', pa3 produced two fluent instances of *v pořádku* 'in order, all right' and one disfluent token of *v pondělí* 'on Monday'; pa2's fluent tokens were *ve vsi* 'in a village', *v zimě* 'in winter', *z konve* 'from a can' and *v pondělí* 'on Monday'.

	clusters 1 and 3				clusters 2 and 4			
	fluent	pre-phrasal	phrase-internal	% fluent	fluent	pre-phrasal	phrase-internal	% fluent
aa1	69	18	11	70.41	17	5	0	77.27
aa3	12	8	2	54.55	14	3	0	82.35
ba1	8	3	5	50	5	4	1	50
ba2	16	12	15	37.21	6	4	2	50
ba3	2	10	3	13.33	3	4	1	37.5
pa1	45	24	4	61.64	7	2	0	77.78

Table 32: Distribution of disfluencies and fluency rate of selected participants with respect to cluster membership.

A comparison of frequency effects with the modified PPs can, effectively, only be made for participants aa1 and aa3. As shown in Figure 52, other participants produced only very few modified PPs from clusters 2 and 4 or no such tokens at all. The three disfluent cluster 2/4 tokens produced by aa3 have a pre-phrasal disfluency. aa1 has a high fluency rate for cluster 2/4 PPs with four pre-phrasal and three phrase-internal disfluencies. The majority of the fluent items in this group are expressions of time that have either a demonstrative modifier (e.g. v tý době 'at that time') or an adjective derived from a numeral used to express a year, such as *do šedesátýho roku* 'til 1960, lit. til the sixtieth year'.

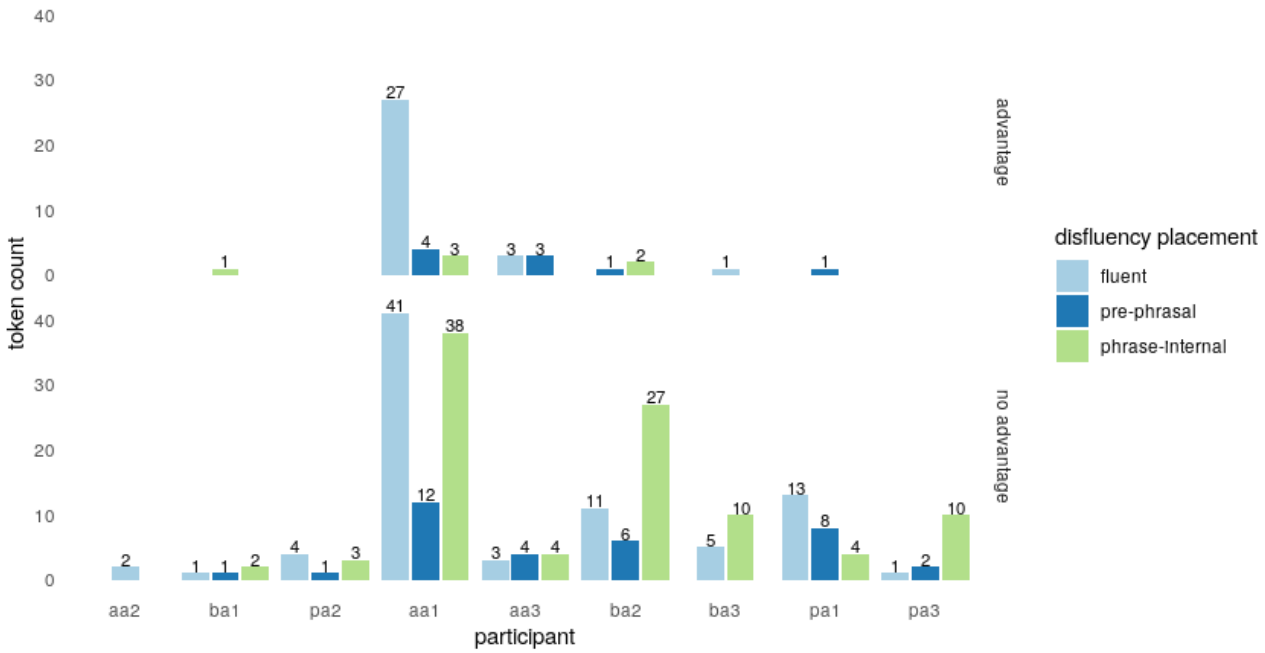

Figure 52: Disfluency placement in modified PPs produced by participants with respect to cluster membership.

While this general trend observed in the data is interesting, it is also possible to focus on the tokens in clusters 1 and 3 and potential differences between those. The frequency index that was described above and used to enhance or complement the cluster analysis can be used for this purpose. While cluster membership is a categorical variable, it is important to bear in mind that the measures per se, used to cluster the data, are indeed continuous and differences between individual tokens cannot be accounted for while using cluster memberships. The scoring index, on the other hand, may be used as an intermediary level of granularity between cluster assignments and individual "configurations" of measures. For this purpose, the PP tokens were labeled with an overall high score index if three or more of the five measures used in the scoring received a high score and no more than one of the measures and, conversely an overall low score label was used if the token received at least three partial low score and no more than a single partial high score. The remaining tokens were classified as mid-range in this procedure. Table 33 shows the distribution of scores between the four clusters for all PPs.

overall score	cluster 1	cluster 2	cluster 3	cluster 4
high	26	16	44	90
low	71	0	25	0
mid-range	210	16	154	11

Table 33: Relationship between cluster membership and score for individual PP tokens.

Before I focus on cluster1/3 unmodified PPs, I would like to discuss the relationship between scores and disfluencies for the whole sample as well as at individual level, similarly to the characteristic using cluster membership. Figure 53 shows this relationship for modified and unmodified PPs for the corpus. We see that the pattern observed in the cluster-based account is visible here as well, it might even be said that it is to an extent clearer. If we look at the unmodified PPs first, the high score tokens are more fluent (67.2 %) and have a lower proportion of phrase-internal disfluencies (19.51 %) than both the mid-score (52.79 % fluent, 33.64 % phrase internal disfluencies) and the low-score (50 % fluent, 34.78 % phrase-internal disfluencies) PPs. The situation is also similar for the modified PPs, where the high score items have a fluency rate of 60.78 % and the mid and low score tokens have fluency rates of 37.97 % and 42 % respectively. We also see a lower proportion of phrase-internal disfluencies in high score PPs and a lower proportion of pre-phrasal disfluencies in the low score group.

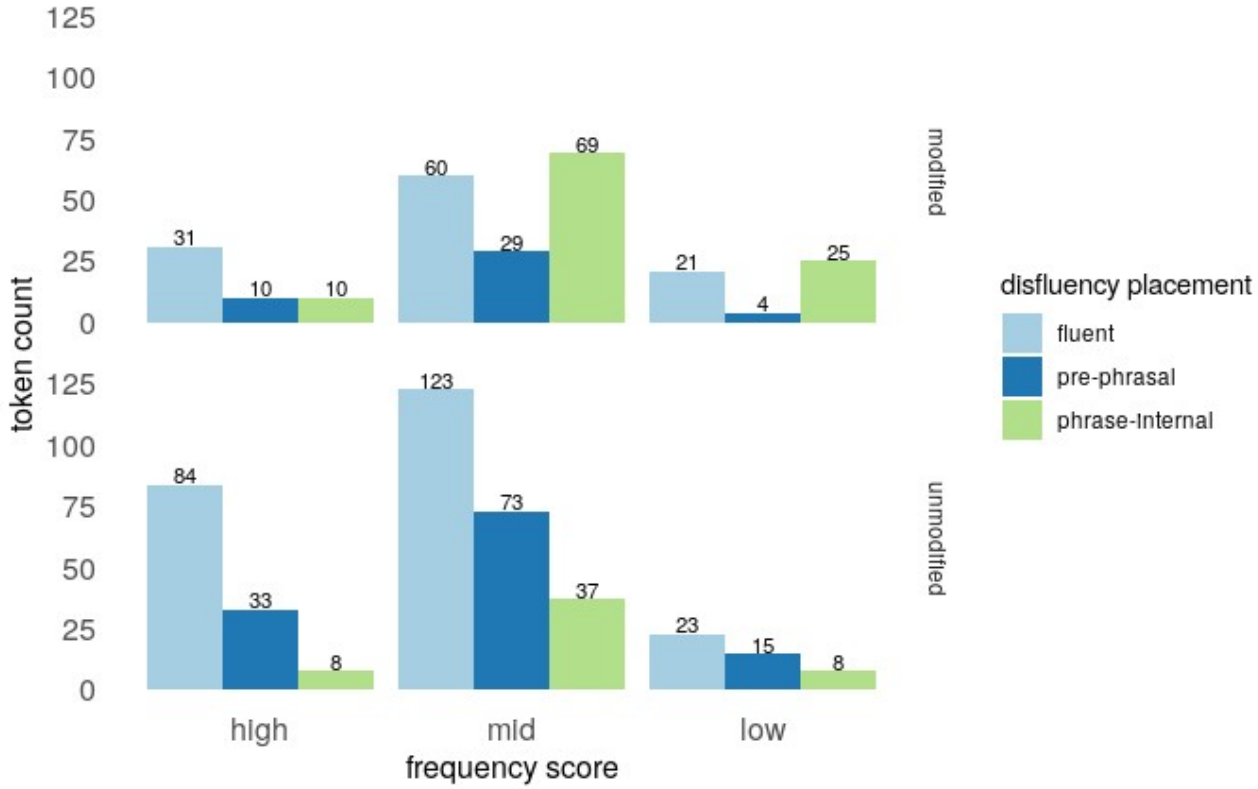

Figure 53: Overall distribution of disfluencies in modified and unmodified PPs with respect to frequency scores.

The performance of individual participants for modified and unmodified tokens together is summarized in Figure 54. The general picture is again very similar to Figure 53. Table 34 shows the distribution of disfluencies across the three scoring categories for participants aa1, aa3, ba2, ba3, and pa1 as a group that produced a number of tokens in all three score ranges and has comparable results when only unmodified PPs are considered.

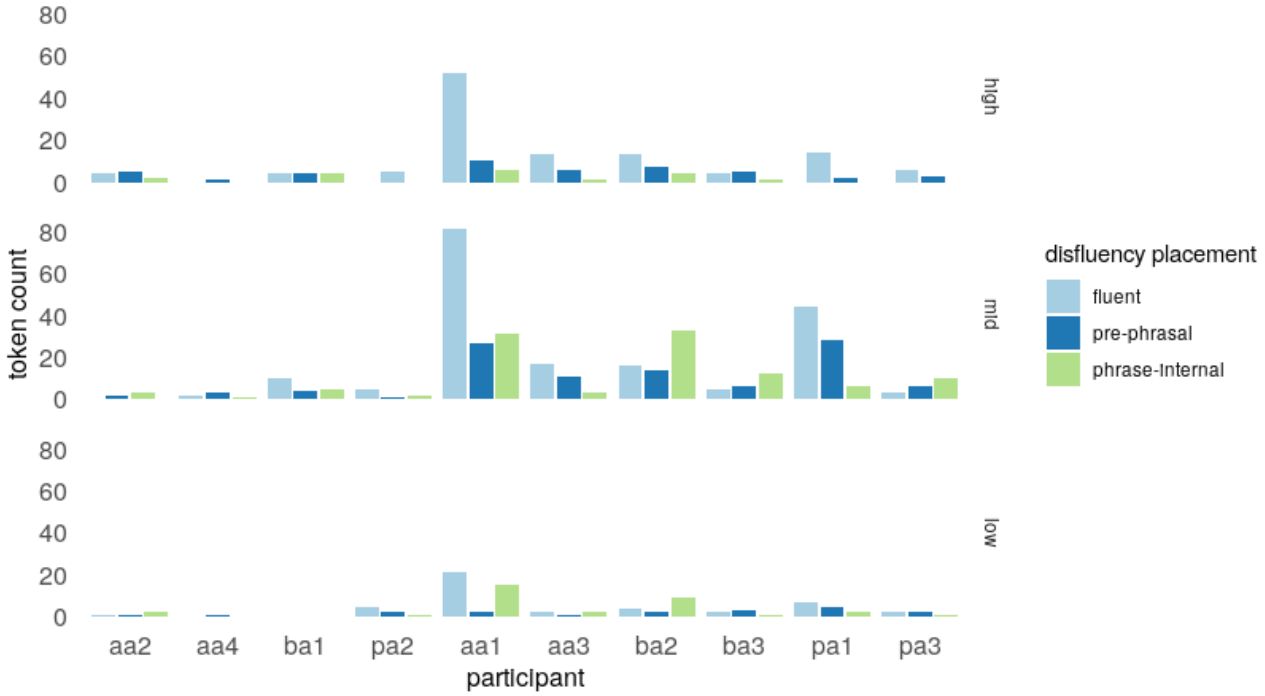

Figure 54: Disfluency placement in all PPs produced by participants with respect to frequency scores.

	high score			mid score			low score		
	fluent	pre-phrasal	phrase-internal	fluent	pre-phrasal	phrase-internal	fluent	pre-phrasal	phrase-internal
aa1	52	10	6	81	27	31	21	2	15
aa3	13	6	1	17	11	3	2	1	2
ba2	13	7	4	16	14	33	4	2	9
ba3	4	5	1	5	6	12	2	3	1
pa1	14	2	0	44	28	6	7	5	2

Table 34: Distribution of disfluencies with regards to overall score for selected participants.

We see again that particularly the high score group has higher fluency rates for all these participants and that phrase-internal disfluencies are less frequent in these tokens. Taking pa1 as an example representing this pattern, we find that 87.5 % of high score PPs were produced fluently and that there are no phrase-internal disfluencies in this group, while the mid and low score tokens have lower fluency levels (56.41 % and 50 % respectively) and that phrase-internal disfluencies occur.

When we look at the disfluency patterns in unmodified cluster 1/3 tokens with regard to these scores, we see again some tendencies in the expected direction, suggesting that some of the remain-

ing variation may be explained by the frequency characteristics of individual tokens. First, when one looks at the partial score counts in these tokens, we find that the mean count of high scores is slightly higher for fluent PPs (1.345) as opposed to pre-phrasal and phrase-internal disfluencies (1.155 and 0.918 respectively) and the reverse is true for the mean number of low scores (0.994 for fluent tokens, 1.052 for pre-phrasal and 1.245 for phrase-internal disfluencies). The distributions of these scores are summarized in 55. The counts of 0 through 3 are of particular interest here as these contain relatively higher numbers of PPs overall. When we look at the upper panel in the figure, we see that tokens with no high score have a lower proportion of fluent productions (45.61 %) and a higher proportion of phrase-internal disfluencies (20.18 %). Fluent production increases with increasing scores (60.23 % for scores of 1, 51.56 % for scores of 2, and 64.71 % for scores of 3) and the proportion of phrase-internal disfluencies simultaneously decreases (14.77 %, 14.06 %, and 5.88 % respectively). The lower panel shows a somewhat reverse pattern in phrase-internal disfluencies that increase from 13.21 % at 0 and 12.31 at 1 to 25.53 % at scores of 2 and 19.05 % at score of 3. However, it should be noted that the pattern is not as clear here as fluency does not markedly decrease with increasing number of low scores.

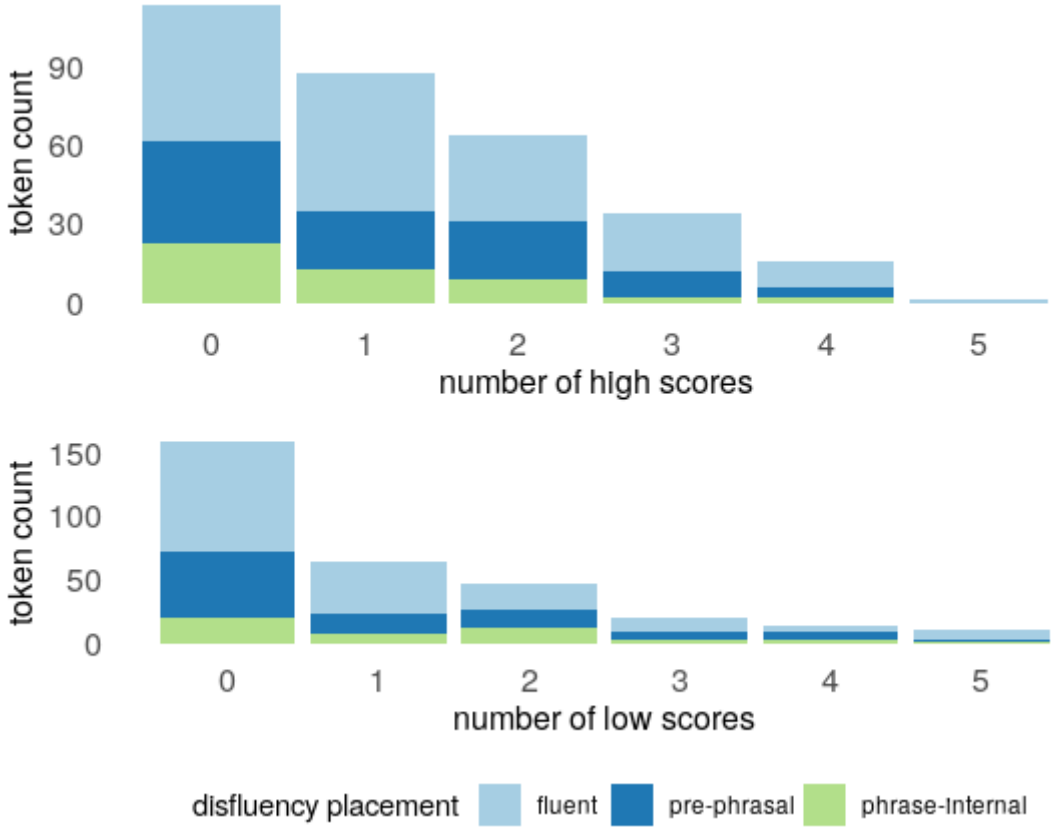

Figure 55: Distribution of disfluencies in clusters 1 and 3 with respect to particular frequency scores.

When we include the high, mid, and low overall scores in the picture, we see a similar pattern of increased fluency in the high score tokens (84 %) accompanied with a lower proportion of phrase-internal disfluencies (22.22 %). Conversely, the low score group has an overall fluency rate of 50 % and the proportion of phrase-internal disfluencies is 34.78 %. These distributions can be seen in

Figure 56, while Figure 57 depicts the same for individual participants. We see that the picture here is not as clear. However, an overall high score seems to play a greater role than low scores here, particularly when we look at the production of participants aa1 and pa1 where we see higher fluency rates for these tokens. All of these suggest that some amount of variation within clusters 1 and 3 is due to the distributional characteristics of the individual PPs.

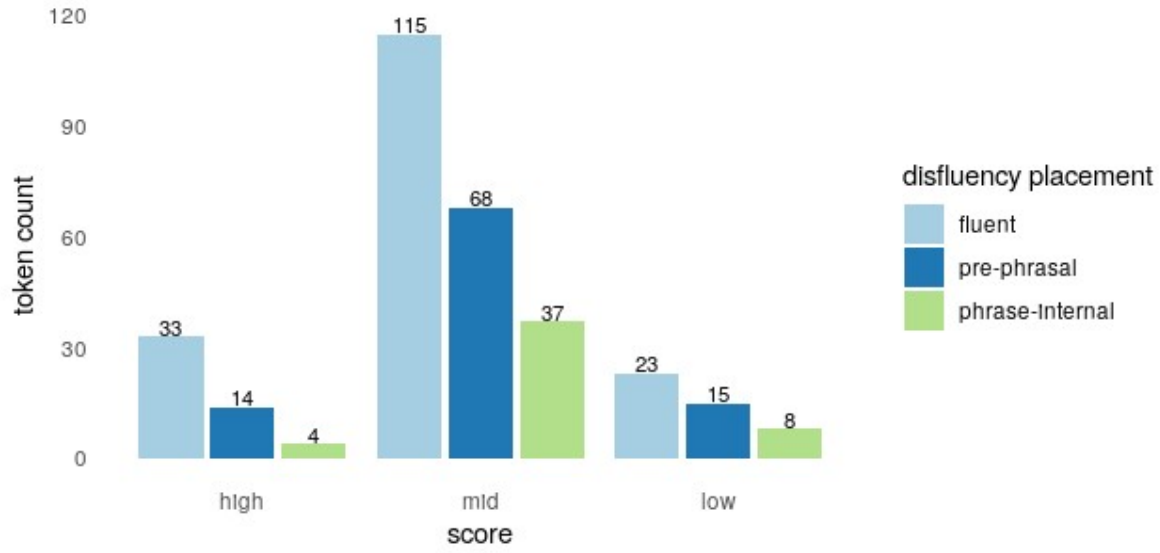

Figure 56: Distribution of disfluencies in unmodified PPs in clusters 1 and 3 with respect to frequency scores.

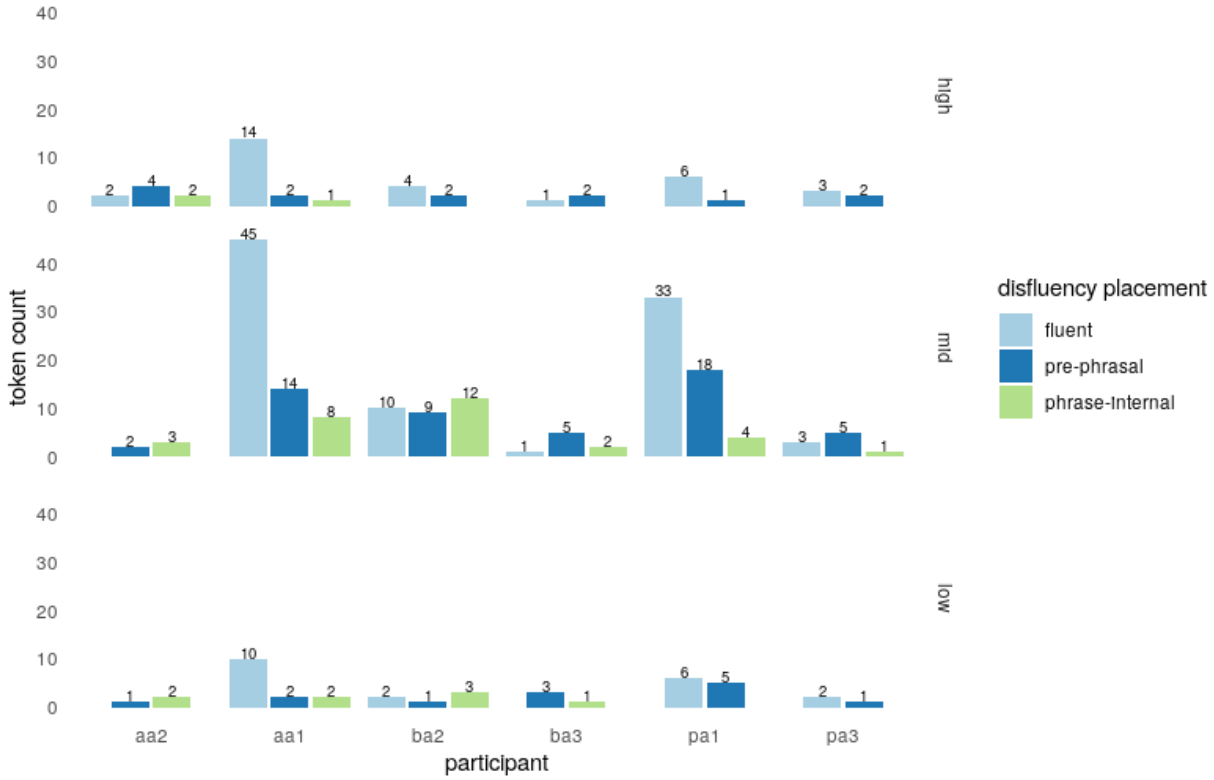

Figure 57: Distribution of disfluencies in unmodified PPs in clusters 1 and 3 produced by individual participants with respect to frequency scores.

The aim of this section was to describe the occurrence of disfluencies in the context structural and usage variables. The results have shown that the syntactic and argument structural status of the PPs in the sample did not influence the fluency of PP production. PPs with prenominal modifiers induced on the whole more disfluent production. This can be due to a greater inner syntactic complexity of the phrases or simply caused by the fact that there are more words in the sequence, since these two factors overlap. There were 19 tokens of postnominally modified complements of which 11 were fluent, all produced by aa1. Five of these tokens are from cluster 2/4. One of the remaining disfluent tokens belonging to cluster 4 had a pre-phrasal disfluency, while the other seven come from clusters 1/3. This is similar to the pattern observed for the completely "bare" PPs, suggesting that it is the length of the PP in words that induces more disfluencies. On the other hand, a similar frequncy-based pattern was also found in the modified phrases and syntactic complexity cannot thus be ruled out.

When it comes to frequency, I suggested a novel approach to assessing its role by integreting distributional data from different corpora and all the potentially relevant measures into a single index of "difficulty" based on the characteristics of the discussed PPs by using unsupervised hiearchical clustering as well as a scoring method grounded in the measures of central tendency. Both of these indices suggest that PPs that contain high frequency complements and combinations of prepositions and nouns that occur frequently in the language are produced more fluently with either no disfluencies or a disfluency before the whole phrase. Under the usage-based model, this would be taken as an indication that such word combinations are more likely directly retrieved from memory as ready-made chunks (Zeschel 2008; Bybee 2010: chap. 3; Schneider 2016). Figure 58 proposes a simple model with four different types of PPs. First, there are collocations with extremely high association measures, represented by *v pohodě* 'alright, lit. in ease' in the figure. These are highly lexicalized expressions whose semantic as well as structural compositionality can be taken to be questionable at best. These expressions can be assumed to be retrieved as chunks more akin to single word units leaving little space for intrusive disfluencies. Secondly, there are case represented in the figure by *v kleci* 'in a/the cage'. The *klec* lemma appears in approximately 30 % of all its uses in this PP. While, the high level of association between the preposition and the noun may also cause a direct retrieval of this bigram. However, the higher compositionality of this PP in comparison to *v pohodě* makes a disfluency slightly more likely to occur phrase-internally, as suggested by the dashed line in the figure. The phrase *k lvovi* 'to a/the lion' represents the next situation where there is no strong association between the preposition and the complement. This is visualized by showing that the PP is composed of a slot for the preposition which is combined with an inflected noun. In more detail, we can conceptualize this such that there is a PP construction with a lexically specific preposition induced, e.g. by an argument structure construction of coming/arriving, and an open slot for a noun in the inflected form governed by the preposition. This PP construction is superimposed with the Inflected noun construction in language production and it is precisely this step that can be assumed to produce more phrase-internal disfluencies. The last schema in the figure shows on the example of *s vařicí vodou* 'with hot water' how an additional step that concerns the combination of the complement noun with a prenominal modifier opens up more space for retrieval problems indexed by disfluencies.

One should keep in mind that the frequency effetcs discussed here must be framed in terms of tendencies rather than clear strong patterns. There are individual differences as well as a large amount

of unexplained variation and additional research is needed, preferably under more controlled conditions. However, this analysis has found a trend qualitatively similar to Schneider (2016)'s study of neurotypical production in English providing additional support for the appeal formulated by Gahl and Menn (2016) who argue for the potential and importance of probabilistic processing and frequency effects above word level in aphasia research. These results also have implications for the assessment and therapy of language in aphasia, as discussed in the concluding remarks.

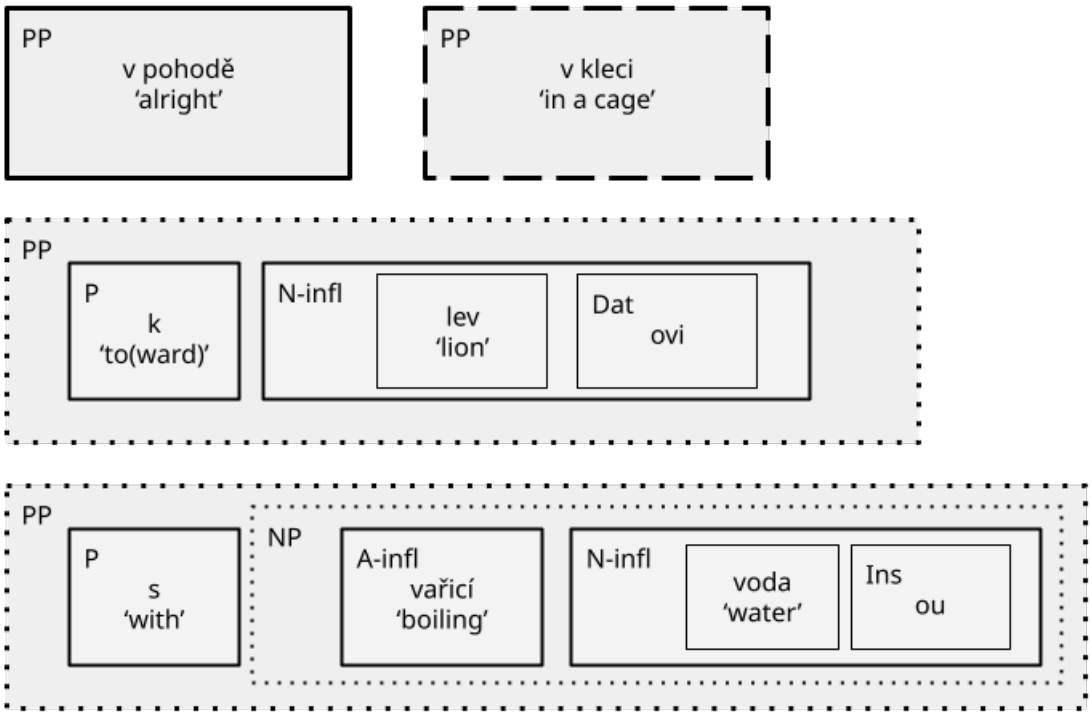

Figure 58: A schema of retrieval of different PP types as modulated by frequency measures.

5 Conclusions

The present work had two main sets of aims. In the first part, I described the process of assembling a corpus of Czech aphasic speech. The corpus consists of samples of descriptive, narrative, procedural, and conversational discourse of eleven speakers with different aphasia types with the total size of almost 17000 words produced by speakers with aphasia. Using the data from the corpus, I performed an analysis of aphasic discourse production. The majority of the microstructural measures selected for this analysis successfully differentiated between neurotypical speakers and participants with non-fluent aphasia. Furthermore, the combination of these measures used to visualize the structure of the sample with hierarchical clustering was able to identify subgroups of participants with similar language profiles. To follow up on these two outcomes, the corpus will be made available for researchers and clinicians.[60] A simple protocol containing the stimulus materials used to collect the data for the corpus as well as a transcription manual and an ELAN template file to enable other specialist to contribute their own data. The measures presented in this work may be con-

60 As of now, both the whole corpus and the subcorpus with detailed annotation are available for specialists in an osf.io repository (TODO). A publication under the LINDAT/CLARIAH-cz infrastructure is planned in the future.

verted into a tool for discourse analysis that could be used to help SLTs in assessment. However, the validation of the measures is needed on a larger sample.

The application of the usage-based framework on the corpus data and the results of the analyses presented in the second part have implications both for linguistics and for the study of aphasia. I hope to have shown the importance of various frequency measures in the analysis of language processing in aphasia. Cumulative lemma frequency was used for the analyses of verb production in the corpus. High frequency was found to increase the probability of fluent production. It was also apparent that persons with more severe lexical retrieval problems rely more on high frequency verbs. There was also a weak correlation between transitivity and frequency in the group of speakers with nonfluent aphasia, showing the importance to account for frequency and similar variables in the analysis of structural phenomena.

The results of the analyses of nominal inflection and disfluency placement in prepositional phrases is of particular interest both for linguistic theory and clinical practice. Both of these analyses suggest that the relative frequencies of particular inflected word forms are related to ease of access and, consequently, success of retrieval in aphasic language production. This is in line with previous research that has shown that the effects of probabilistic processing are present in both neurotypical speakers and persons with aphasia and the differences between the groups are qualitative rather than quantitative in nature. The fact that cooccurrence frequency and association strength between particular prepositions and nominal complements can be associated to fluent or disfluent production of prepositional phrases provides additional evidence for the usage-based model. Turning to clinical practice, the present work provides arguments for a tighter cooperation between linguists and clinicians and, in particular, for the introduction of corpus linguistic data in clinical practice. If we take the patterns observed in the analysis of nouns, these have the potential to be directly applied in SLT. When working with a patient in therapy on nominal morphology in Czech, a SLT can focus on specific high frequency word forms in typical usage contexts rather than using more standard protocols that focus on citation forms or whole paradigms.

In conclusion, I hope to have contributed to the growing body of literature that shows that the usage-based framework has potential benefits for the study of language in aphasia.The present study has several limitations and follow up studies with more participants and more controlled elicitation are necessary. However, my objective was to demonstrate the potential of the usage-based framework to discover interesting trends and patterns in the data that generate research questions and, if corroborated, have important clinical implications. It is precisely these new avenues of research that are subjectively the most important contribution of the present work.

6 References

Abbot-Smith, Kirsten & Michael Tomasello. 2006. Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review* 23(3). https://doi.org/10.1515/TLR.2006.011.

Akhutina, Tatiana. 2016. Luria's classification of aphasias and its theoretical basis. *Aphasiology* 30(8). 878–897. https://doi.org/10/gfkm8b.

Allen, Micah, Davide Poggiali, Kirstie Whitaker, Tom Rhys Marshall, Jordy van Langen & Rogier A. Kievit. 2021. Raincloud plots: a multi-platform tool for robust data visualization [version 2; peer review: 2 approved]. *Wellcome Open Research*. https://doi.org/10.12688/wellcomeopenres.15191.2.

Anderson, Elizabeth C. 2017. *Frequency effects in the processing of verbs and argument structure: Evidence from adults with and without acquired aphasia*. University of Sheffield phd. https://etheses.whiterose.ac.uk/20867/. (5 August, 2021).

Andreetta, Sara, Anna Cantagallo & Andrea Marini. 2012. Narrative discourse in anomic aphasia. *Neuropsychologia* 50(8). 1787–1793. https://doi.org/10.1016/j.neuropsychologia.2012.04.003.

Ardila, Alfredo. 2010. A proposed reinterpretation and reclassification of aphasic syndromes. *Aphasiology* 24(3). 363–394. https://doi.org/10.1080/02687030802553704.

Armstrong, Elizabeth. 2000. Aphasic discourse analysis: The story so far. *Aphasiology* 14(9). 875–892. https://doi.org/10.1080/02687030050127685.

Arnon, Inbal & Uriel Cohen Priva. 2013. More than Words: The Effect of Multi-word Frequency and Constituency on Phonetic Duration. *Language and Speech* 56(3). 349–371. https://doi.org/10.1177/0023830913484891.

Avrutin, Sergey. 2001. Linguistics and agrammatism. *Glot International* 5(3). 87–97.

Baayen, Harald R., James M. McQueen, Ton Dijkstra & Robert Schreuder. 2003. Frequency effects in regular inflectional morphology: Revisiting Dutch plurals. In R Harald Baayen & Robert Schreuder (eds.), *Morphological Structure in Language Processing* (Trends in Linguistics : Studies and Monographs 151), 355–390. Mouton de Gruyter.

Baayen, R. Harald. 2009. Corpus linguistics in morphology: morphological productivity. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. An international handbook, Vol. 2*, 899–919. Berlin: De Gruyter Mouton.

Badecker, William & Alfonso Caramazza. 1985. On considerations of method and theory governing the use of clinical categories in neurolinguistics and cognitive neuropsychology: The case against agrammatism. *Cognition* 20(2). 97–125. https://doi.org/10.1016/0010-0277(85)90049-6.

Barking, Marie, Ad Backus & Maria Mos. 2022. Individual corpus data predict variation in judgments: testing the usage-based nature of mental representations in a language transfer setting. *Cognitive Linguistics*. De Gruyter Mouton. https://doi.org/10.1515/cog-2021-0105.

Bastiaanse, Roelien, Gosse Bouma & Wendy Post. 2009. Linguistic complexity and frequency in agrammatic speech production. *Brain and Language* 109(1). 18–28. https://doi.org/10.1016/j.bandl.2008.12.004.

Bastiaanse, Roelien & Roel Jonkers. 2012. Linguistic accounts of agrammatic aphasia. In Roelien Bastiaanse & Cynthia K Thompson (eds.), *Perspectives on agrammatism*, 17–33. Psychology Press.

Bastiaanse, Roelien, Martijn Wieling & Nienke Wolthuis. 2016. The role of frequency in the retrieval of nouns and verbs in aphasia. *Aphasiology* 30(11). 1221–1239. https://doi.org/10.1080/02687038.2015.1100709.

Bastiaanse, Roelien & Ron van Zonneveld. 2005. Sentence production with verbs of alternating transitivity in agrammatic Broca's aphasia. *Journal of Neurolinguistics* 18(1). 57–66. https://doi.org/10.1016/j.jneuroling.2004.11.006.

Bates, Elizabeth, Angela Friederici & Beverly Wulfeck. 1987. Comprehension in aphasia: A cross-linguistic study. *Brain and language* 32(1). 19–67.

Bates, Elizabeth & Judith C. Goodman. 1997. On the Inseparability of Grammar and the Lexicon: Evidence from Acquisition, Aphasia and Real-time Processing. *Language and Cognitive Processes* 12(5–6). 507–584. https://doi.org/10.1080/016909697386628.

Bates, Elizabeth & Brian MacWhinney. 1987. Competition, variation, and language learning. In Brian MacWhinney (ed.), *Mechanisms of language aquisition*, 157–193. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bates, Elizabeth & Brian MacWhinney. 1989. Functionalism and the competition model. In *The crosslinguistic study of sentence processing*, vol. 3, 73–112.

Beckner, Clay, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman & Tom Schoenemann. 2009. Language Is a Complex Adaptive System: Position Paper. *Language Learning* 59. 1–26. https://doi.org/10.1111/j.1467-9922.2009.00533.x.

Benson, D F & Norman Geschwind. 1971. Aphasia and related cortical disturbances. In A B Baker & L H Baker (eds.), *Clinical Neurology*. Harper & Row.

Benton, Arthur & Steven W. Anderson. 1998. Aphasia: Historical Perspectives. In Martha Taylor Sarno (ed.), *Acquired Aphasia*, 1–24. 3rd edn. Elsevier. https://doi.org/10.1016/B978-012619322-0/50004-X.

Benton, Arthur & Robert J. Joynt. 1960. Early Descriptions of Aphasia. *Archives of Neurology* 3(2). 205–222. https://doi.org/10/cx39jg.

Bergelson, Mira B. & Mariya V. Khudyakova. 2020. *Chapter 8. Narrative discourse segmentation in clinical linguistics* (Studies in Corpus Linguistics). (Ed.) Shlomo Izre'el, Heliana Mello, Alessandro Panunzi & Tommaso Raso. *In Search of Basic Units of Spoken Language: A corpus-driven approach* (Studies in Corpus Linguistics). John Benjamins Publishing Company. https://doi.org/10.1075/scl.94.08ber.

Berndt, Rita Sloan, Charlotte C. Mitchum & Anne N. Haendiges. 1996. Comprehension of reversible sentences in "agrammatism": a meta-analysis. *Cognition* 58(3). 289–308. https://doi.org/10.1016/0010-0277(95)00682-6.

Berndt, Rita Sloan, Myrna Schwarz & Eleanor Saffran. 2001. *Quantitative Production Analysis: A Training Manual for the Analysis of Aphasic Sentence Production*. Psychology Press.

Berthier, Marcelo L. 2005. Poststroke aphasia : epidemiology, pathophysiology and treatment. *Drugs & Aging* 22(2). 163–182. https://doi.org/10/dkhfkx.

Beveridge, Madeleine E. L. & Thomas H. Bak. 2011. The languages of aphasia research: Bias and diversity. *Aphasiology* 25(12). 1451–1468. https://doi.org/10.1080/02687038.2011.624165.

Beyersmann, Elisabeth, Eleanor M. Dutton, Sohaila Amer, Niels O. Schiller & Britta Biedermann. 2015. The production of singular- and plural-dominant nouns in Dutch. *Language, Cognition and Neuroscience*. Routledge 30(7). 867–876. https://doi.org/10.1080/23273798.2015.1027236.

Biedermann, Britta, Elisabeth Beyersmann, Catherine Mason, Franziska Machleb, Mareike Moormann & Antje Lorenz. 2018. Production of German –n plurals in aphasia: Effects of dominance and predictability. *Aphasiology* 32(9). 1087–1112. https://doi.org/10.1080/02687038.2017.1387228.

Biedermann, Britta, Antje Lorenz, Elisabeth Beyersmann & Lyndsey Nickels. 2012. The influence of plural dominance in aphasic word production. *Aphasiology* 26(8). 985–1004. https://doi.org/10.1080/02687038.2012.660459.

Blackwell, Arshavir & Elizabeth Bates. 1995. Inducing Agrammatic Profiles in Normals: Evidence for the Selective Vulnerability of Morphology under Cognitive Resource Limitation. *Journal of Cognitive Neuroscience* 7(2). 228–257. https://doi.org/10.1162/jocn.1995.7.2.228.
Bod, Rens. 2006. Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review* 23(3). https://doi.org/10.1515/TLR.2006.012.
Bod, Rens, Jennifer Hay & Stefanie Jannedy (eds.). 2003. *Probabilistic linguistics*. Cambridge, MA: MIT Press.
Boles, Larry. 1998. Conversational discourse analysis as a method for evaluating progress in aphasia: a case report. *Journal of Communication Disorders* 31(3). 261–274. https://doi.org/10.1016/S0021-9924(98)00005-7.
Booij, Geert. 2010a. Construction Morphology. *Language and Linguistics Compass* 4(7). 543–555. https://doi.org/10.1111/j.1749-818X.2010.00213.x.
Booij, Geert. 2010b. *Construction Morphology*. Oxford, New York: Oxford University Press.
Booij, Geert. 2018. *The construction of words, advances in construction morphology*. New York, NY: Springer Berlin Heidelberg.
Boye, Kasper & Roelien Bastiaanse. 2018. Grammatical versus lexical words in theory and aphasia: Integrating linguistics and neurolinguistics. *Glossa: a journal of general linguistics* 3(1). 29. https://doi.org/10.5334/gjgl.436.
Boye, Kasper & Peter Harder. 2012. A usage-based theory of grammatical status and grammaticalization. *Language* 88(1). 1–44. https://doi.org/10.1353/lan.2012.0020.
Brown, Cati, Tony Snodgrass, Susan J. Kemper, Ruth Herman & Michael A. Covington. 2008. Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods* 40(2). 540–545. https://doi.org/10.3758/BRM.40.2.540.
Bruns, Claudia, Suzanne Beeke, Vitor C. Zimmerer, Carolyn Bruce & Rosemary A. Varley. 2021. Training flexibility in fixed expressions in non-fluent aphasia: A case series report. *International Journal of Language & Communication Disorders* 56(5). 1009–1025. https://doi.org/10.1111/1460-6984.12652.
Bryant, Lucy, Alison Ferguson & Elizabeth Spencer. 2016. Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics & Phonetics* 30(7). 489–518. https://doi.org/10.3109/02699206.2016.1145740.
Bryant, Lucy, Elizabeth Spencer & Alison Ferguson. 2017. Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology* 31(10). 1105–1126. https://doi.org/10.1080/02687038.2016.1239013.
Brysbaert, Marc & Andrew W. Ellis. 2016. Aphasia and age of acquisition: are early-learned words more resilient? *Aphasiology* 30(11). 1240–1263. https://doi.org/10.1080/02687038.2015.1106439.
Brysbaert, Marc & Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41(4). 977–990. https://doi.org/10.3758/BRM.41.4.977.
Bybee, Joan. 2010. *Language, Usage and Cognition.* Leiden: Cambridge University Press. http://public.eblib.com/choice/publicfullrecord.aspx?p=534754. (15 January, 2019).
Bybee, Joan L. 2006. From Usage to Grammar: The Mind's Response to Repetition. *Language* 82(4). 711–733. https://doi.org/10.1353/lan.2006.0186.
Bybee, Joan L. 2007. *Frequency of use and the organization of language*. Oxford ; New York: Oxford University Press.
Bybee, Joan & James L. McClelland. 2005. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review* 22(2–4). https://doi.org/10.1515/tlir.2005.22.2-4.381.

Campione, Estelle & Jean Véronis. 2002. A Large-Scale Multilingual Study of Silent Pause Duration. In.
Caramazza, Alfonso & Michael McCloskey. 1988. The case for single-patient studies. *Cognitive Neuropsychology*. Routledge 5(5). 517–527. https://doi.org/10.1080/02643298808253271.
Caramazza, Alfonso & Edgar B. Zurif. 1976. Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language* 3(4). 572–582. https://doi.org/10.1016/0093-934X(76)90048-1.
Carota, A., A. O. Rossetti, T. Karapanayiotides & J. Bogousslavsky. 2001. Catastrophic reaction in acute stroke: A reflex behavior in aphasic patients. *Neurology* 57(10). 1902–1905. https://doi.org/10.1212/WNL.57.10.1902.
Casilio, Marianne, Kindle Rising, Pélagie M. Beeson, Kate Bunton & Stephen M. Wilson. 2019. Auditory-Perceptual Rating of Connected Speech in Aphasia. *American Journal of Speech-Language Pathology*. American Speech-Language-Hearing Association 28(2). 550–568. https://doi.org/10.1044/2018_AJSLP-18-0192.
Chater, Nick & Morten H. Christiansen. 2016. Squeezing through the Now-or-Never bottleneck: Reconnecting language processing, acquisition, change, and structure. *Behavioral and Brain Sciences* 39. https://doi.org/10.1017/S0140525X15001235.
Cho, Soojin & Cynthia K. Thompson. 2010. What goes wrong during passive sentence production in agrammatic aphasia: An eyetracking study. *Aphasiology* 24(12). 1576–1592. https://doi.org/10.1080/02687031003714442.
Choy, Jungwon Janet & Cynthia K. Thompson. 2010. Binding in agrammatic aphasia: Processing to comprehension. *Aphasiology*. Routledge 24(5). 551–579. https://doi.org/10.1080/02687030802634025.
Christiansen, Morten H. & Nick Chater. 2015. The language faculty that wasn't: a usage-based account of natural language recursion. *Frontiers in Psychology* 6. https://doi.org/10.3389/fpsyg.2015.01182.
Christiansen, Morten H. & Nick Chater. 2016. The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences* 39. e62. https://doi.org/10.1017/S0140525X1500031X.
Chromý, Jan, Vojtěch Diatka & Olga Džupová. 2015. Kategoriální normy češtiny pro 12 kategorií. *Studie z aplikované lingvistiky* 6(2). 26–51.
Code, Chris. 2013. Significant Landmarks in the History of Aphasia and its Therapy. In Ilias Papathanasiou, Patrick Coppens & Constantin Potagas (eds.), *Aphasia and related neurogenic communication disorders*, 3–22. Burlington, MA: Jones & Bartlett Learning.
Code, Chris & Brian Petheram. 2011. Delivering for aphasia. *International Journal of Speech-Language Pathology*. Taylor & Francis 13(1). 3–10. https://doi.org/10.3109/17549507.2010.520090.
Coelho, Carl, Karen Lê, Jennifer Mozeiko, Frank Krueger & Jordan Grafman. 2012. Discourse production following injury to the dorsolateral prefrontal cortex. *Neuropsychologia* 50(14). 3564–3572. https://doi.org/10.1016/j.neuropsychologia.2012.09.005.
Croft, William A. 2001. *Radical Construction Grammar*. Oxford University Press.
Cséfalvay, Zsolt, Marta Klimešová & Milena Košťálová. 2003. *Diagnostika a terapie afázie, alexie, agrafie: (manuál)*. Praha: Asociace klinických logopedů.
Cvrček, Václav. 2019. Calc: Korpusová kalkulačka. Ústav Českého národního korpusu FF UK. https://www.korpus.cz/calc/. (28 August, 2022).
Cvrček, Václav & Lucie Chlumská. 2015. Simplification in translated Czech: a new approach to type-token ratio. *Russian Linguistics* 39(3). 309–325. https://doi.org/10/gg3hdm.
Dąbrowska, Ewa. 2004. Rules or schemas? Evidence from Polish. *Language and Cognitive Processes* 19(2). 225–271. https://doi.org/10.1080/01690960344000170.

Dąbrowska, Ewa. 2010. Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27(1). 1–23. https://doi.org/10.1515/tlir.2010.001.
Dąbrowska, Ewa. 2012. Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism* 2(3). 219–253. https://doi.org/10.1075/lab.2.3.01dab.
Dąbrowska, Ewa & Elena Lieven. 2005. Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics* 16(3). https://doi.org/10.1515/cogl.2005.16.3.437.
Damasio, Hanna. 2008. Neural Basis of Language Disorders. In Roberta Chapey (ed.), *Language Intervention Strategies in Aphasia and Related Neurogenic Communication Disorders*, 40–61. 5th edn. Lippincott Williams & Wilkins.
Damico, Jack S., Mary Oelschlaeger & Nina Simmons-Mackie. 1999. Qualitative methods in aphasia research: conversation analysis. *Aphasiology* 13(9–11). 667–679. https://doi.org/10.1080/026870399401777.
DeDe, Gayle. 2012. Lexical and Prosodic Effects on Syntactic Ambiguity Resolution in Aphasia. *Journal of Psycholinguistic Research* 41(5). 387–408. https://doi.org/10.1007/s10936-011-9191-1.
DeDe, Gayle. 2013a. Verb transitivity bias affects on-line sentence reading in people with aphasia. *Aphasiology* 27(3). 326–343. https://doi.org/10.1080/02687038.2012.725243.
DeDe, Gayle. 2013b. Effects of verb bias and syntactic ambiguity on reading in people with aphasia. *Aphasiology* 27(12). 1408–1425. https://doi.org/10.1080/02687038.2013.843151.
Dell, Gary S., Myrna F. Schwartz, Nadine Martin, Eleanor M. Saffran & Deborah A. Gagnon. 1997. Lexical access in aphasic and nonaphasic speakers. *Psychological Review* 104(4). 801–838. https://doi.org/10.1037/0033-295X.104.4.801.
Dickey, Michael Walsh, JungWon Janet Choy & Cynthia K. Thompson. 2007. Real-time comprehension of wh- movement in aphasia: Evidence from eyetracking while listening. *Brain and Language* 100(1). 1–22. https://doi.org/10.1016/j.bandl.2006.06.004.
Diessel, Holger. 2015. Usage-based construction grammar. In Ewa Dabrowska & Dagmar Divjak (eds.), *Handbook of Cognitive Linguistics*. Berlin, München, Boston: DE GRUYTER. https://doi.org/10.1515/9783110292022-015.
Diessel, Holger. 2020. A Dynamic Network Approach to the Study of Syntax. *Frontiers in Psychology*. Frontiers 11. https://doi.org/10/gh3s2g.
DiLallo, Jennifer, Heidi Mettler & Gayle DeDe. 2017. Corpus-based transitivity biases in individuals with aphasia. *Aphasiology* 31(4). 447–464. https://doi.org/10.1080/02687038.2016.1271105.
Divjak, Dagmar. 2019. *Frequency in Language: Memory, Attention and Learning*. 1st edn. Cambridge University Press. https://doi.org/10.1017/9781316084410.
Divjak, Dagmar & Antti Arppe. 2013. Extracting prototypes from exemplars What can corpus data tell us about concept representation? *Cognitive Linguistics* 24(2). https://doi.org/10.1515/cog-2013-0008.
Divjak, Dagmar, Ewa Dąbrowska & Antti Arppe. 2016. Machine Meets Man: Evaluating the Psychological Reality of Corpus-based Probabilistic Models. *Cognitive Linguistics* 27(1). 1–33. https://doi.org/10.1515/cog-2015-0101.
Divjak, Dagmar, Natalia Levshina & Jane Klavan. 2016. "Cognitive Linguistics: Looking back, looking forward." *Cognitive Linguistics* 27(4). 447–463. https://doi.org/10.1515/cog-2016-0095.
Eling, Paul (ed.). 1994. *Reader in the history of aphasia: from Franz Gall to Norman Geschwind* (Amsterdam Studies in the Theory and History of Linguistic Science v. 4). Amsterdam ; Philadelphia: J. Benjamins Pub. Co.

Ellis, Charles & Maude Rittman. 2009. Disfluency: An Exploratory Study of the Effects of Subcortical Stroke. *Contemporary Issues in Communication Science and Disorders* 36. 149–156.
Feyereisen, Pierre, Charlotte Verbeke-dewitte & Xavier Seron. 1986. On fluency measures in aphasic speech. *Journal of Clinical and Experimental Neuropsychology* 8(4). 393–404. https://doi.org/10.1080/01688638608401329.
Fillmore, Charles J. 1988. The Mechanisms of "Construction Grammar." In *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*.
Flanderková, Eva. 2019. *Čeština v afázii: teorie a empirie*. Karolinum.
Flanderková, Eva, Barbara Mertins, Ondřej Bezdíček, Eva Baborová & M. Černá. 2014. Posuzování gramatičnosti v Brocově afázii - příklad dvou pacientů. *Česká a slovenská neurologie a neurochirurgie* 77/110(2). 202–209.
Fors, Kristina Lundholm. 2011. An investigation of intra-turn pauses in spontaneous speech. In *Proceedings from Fonetik 2011, Speech, Music and Hearing, KTH, Stockholm, TMH-QPSR*, vol. 51, 65–68. Department of Speech, Music and Hearing, KTH, Stockholm.
Fraundorf, Scott H. & Duane G. Watson. 2014. Alice's adventures in um-derland: psycholinguistic sources of variation in disfluency production. *Language, Cognition and Neuroscience* 29(9). 1083–1096. https://doi.org/10.1080/01690965.2013.832785.
Fried, Mirjam. 2013. Pojem konstrukce v konstrukční gramatice. *Časopis pro moderní filologii* 95(1). 9–27.
Fried, Mirjam & Jan-Ola Östman (eds.). 2004. *Construction grammar in a cross-language perspective* (Constructional Approaches to Language v. 2). Amsterdam ; Philadelphia: John Benjamins Pub.
Friedmann, Na'ama. 2001. Agrammatism and the Psychological Reality of the Syntactic Tree. *Journal of Psycholinguistic Research* 30(1). 71–90.
Friedmann, Naama. 2006. Speech Production in Broca's Agrammatic Aphasia: Syntactic Tree Pruning. In Yosef Grodzinsky & Katrin Amunts (eds.), *Broca's Region*, 63–82. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195177640.003.0005.
Friedmann, Na'ama & Yosef Grodzinsky. 1997. Tense and Agreement in Agrammatic Production: Pruning the Syntactic Tree. *Brain and Language* 56(3). 397–425. https://doi.org/10.1006/brln.1997.1795.
Fromm, Davida, Margaret Forbes, Audrey Holland, Sarah Grace Dalton, Jessica Richardson & Brian MacWhinney. 2017. Discourse Characteristics in Aphasia Beyond the Western Aphasia Battery Cutoff. *American Journal of Speech-Language Pathology* 26(3). 762–768. https://doi.org/10.1044/2016_AJSLP-16-0071.
Fromm, Davida, Joel Greenhouse, Mitchell Pudil, Yichun Shi & Brian MacWhinney. 2021. Enhancing the classification of aphasia: a statistical analysis using connected speech. *Aphasiology*. Routledge 0(0). 1–28. https://doi.org/10/gmv6nn.
Gahl, Susanne. 2000. *A usage-based model of aphasic sentence comprehension*. University of California Berkeley Dissertation.
Gahl, Susanne. 2002. Lexical biases in aphasic sentence comprehension: An experimental and corpus linguistic study. *Aphasiology* 16(12). 1173–1198. https://doi.org/10.1080/02687030244000428.
Gahl, Susanne & Lise Menn. 2016. Usage-based approaches to aphasia. *Aphasiology* 30(11). 1361–1377. https://doi.org/10.1080/02687038.2016.1140120.
Gahl, Susanne, Lise Menn, Gail Ramsberger, Daniel S. Jurafsky, Elizabeth Elder, Molly Rewega & L. Holland Audrey. 2003. Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences. *Brain and Cognition* 53(2). 223–228. https://doi.org/10.1016/S0278-2626(03)00114-3.

Garnsey, Susan M., Neal J. Pearlmutter, Elizabeth Myers & Melanie A. Lotocky. 1997. The Contributions of Verb Bias and Plausibility to the Comprehension of Temporarily Ambiguous Sentences. *Journal of Memory and Language* 37(1). 58–93. https://doi.org/10.1006/jmla.1997.2512.
Gibbs, Raymond W. 2006. Introspection and cognitive linguistics: Should we trust our own intuitions? *Annual Review of Cognitive Linguistics*. John Benjamins 4(1). 135–151. https://doi.org/10.1075/arcl.4.06gib.
Gibson, Edward, Chaleece Sandberg, Evelina Fedorenko, Leon Bergen & Swathi Kiran. 2016. A rational inference approach to aphasic language comprehension. *Aphasiology* 30(11). 1341–1360. https://doi.org/10.1080/02687038.2015.1111994.
Gilquin, Gaëtanelle (ed.). 2013. *Errors and disfluencies in spoken corpora* (Benjamins Current Topics Volume 52). Amsterdam ; Philadelphia: John Benjamins Pub. Company.
Glosser, Guila & Toni Deser. 1991. Patterns of discourse production among neurological patients with fluent language disorders. *Brain and Language* 40(1). 67–88. https://doi.org/10.1016/0093-934X(91)90117-J.
Goldberg, Adele E. 1995. *Constructions: a construction grammar approach to argument structure* (Cognitive Theory of Language and Culture). Chicago: University of Chicago Press.
Goldberg, Adele E. 2019. *Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions*. Princeton, NJ: Princeton University Press.
Goldman-Eisler, Frieda. 1958. The Predictability of Words in Context and the Length of Pauses in Speech. *Language and Speech* 1(3). 226–231. https://doi.org/10.1177/002383095800100308.
Goldman-Eisler, Frieda. 1961. The Predictability of Words in Context and the Length of Pauses in Speech. *Journal of Communication* 11(2). 95–99. https://doi.org/10.1111/j.1460-2466.1961.tb00334.x.
Goldman-Eisler, Frieda. 1964. Hesitation, Information, and Levels of Speech Production. In A. V .S. de Reuck & Maeve O'Connor (eds.), *Disorders of Language*, 96–114. J. & A. Churchill. https://doi.org/10.1002/9780470715321.ch6.
Goodglass, Harold & E Kaplan. 1972a. *The assessment of aphasia and related disorders*. Lea & Febiger.
Goodglass, Harold & E Kaplan. 1972b. *Boston Diagnostic Aphasia Examination (BDAE)*. Philadelphia: Lea & Febiger.
Goodglass, Harold & Lise Menn. 1985. Is Agrammatism a Unitary Phenomenon? In Mary-Louise Kean (ed.), *Agrammatism*, 1–26. Elsevier. https://doi.org/10.1016/B978-0-12-402830-2.50005-5.
Grodzinsky, Yosef. 1984. The syntactic characterization of agrammatism. *Cognition* 16(2). 99–120. https://doi.org/10/c4wcjq.
Grodzinsky, Yosef. 1986. Language deficits and the theory of syntax. *Brain and Language* 27(1). 135–159. https://doi.org/10.1016/0093-934X(86)90009-X.
Grodzinsky, Yosef. 1995. A Restrictive Theory of Agrammatic Comprehension. *Brain and Language* 50. 27–51.
Grodzinsky, Yosef. 2000. The Trace Deletion Hypothesis and the Tree-Pruning Hypothesis: Still valid characterizations of Broca's aphasia. *Behavioral and Brain Sciences* 23(1). 55–64. https://doi.org/10.1017/S0140525X00582399.
Grodzinsky, Yosef. 2006. A Blueprint for a Brain Map of Syntax. In Yosef Grodzinsky & Katrin Amunts (eds.), *Broca's Region*, 83–107. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195177640.003.0006.
Hagoort, Peter. 2013. MUC (Memory, Unification, Control) and beyond. *Frontiers in Psychology* 4. https://doi.org/10.3389/fpsyg.2013.00416.

Halliday, M. A. K. & Christian M. I. M. Matthiessen. 2014. *Halliday's Introduction to Functional Grammar*. 4th edn. Routledge. https://www.jstor.org/stable/414947?origin=crossref. (15 January, 2019).
Hallowell, Brooke & Roberta Chapey. 2008. Introduction to Language Intervention Strategies in Adult Aphasia. In Roberta Chapey (ed.), *Language Intervention Strategies in Aphasia and Related Neurogenic Communication Disorders*, 3–19. 5th edn. Baltimore, MD: Lippincott Williams & Wilkins.
Hartsuiker, Robert J. & Herman H.J. Kolk. 1998. Syntactic Facilitation in Agrammatic Sentence Production. *Brain and Language* 62(2). 221–254. https://doi.org/10.1006/brln.1997.1905.
Haspelmath, Martin. 2006. Against markedness (and what to replace it with). *Journal of Linguistics* 42(01). 25. https://doi.org/10.1017/S0022226705003683.
Hatchard, Rachel. 2021. *Cognitive Aphasiology – A Usage-Based Approach to Language in Aphasia* (Constructional Approaches to Language). Vol. 31. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/cal.31.
Hatchard, Rachel & Elena Lieven. 2019. Inflection of nouns for grammatical number in spoken narratives by people with aphasia: how glass slippers challenge the rule-based approach. *Language and Cognition* 11(3). 341–372. https://doi.org/10.1017/langcog.2019.21.
Heeschen, Claus. 1985. Agrammatism versus Paragrammatism: A Fictitious Opposition. In Mary-Louise Kean (ed.), *Agrammatism*, 207–248. Academic Press. https://doi.org/10.1016/B978-0-12-402830-2.50012-2.
Hickok, Gregory & Ursula Bellugi. 2001. The signs of aphasia. In Rita Sloan Berndt (ed.), *Handbook of Neuropsychology*, vol. 3, 31–50. 2nd edn. Elsevier Science.
Himmelmann, Nikolaus P. 2014. Asymmetries in the prosodic phrasing of function words: Another look at the suffixing preference. *Language* 90(4). 927–960. https://doi.org/10.1353/lan.2014.0105.
Howes, Davis. 1964. Application of the Word-Frequency Concept to Aphasia. In A. V .S. de Reuck & Maeve O'Connor (eds.), *Disorders of Language*, 47–78. J. & A. Churchill. https://doi.org/10.1002/9780470715321.ch4.
Howes, Davis H. & R. L. Solomon. 1951. Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology*. US: American Psychological Association 41(6). 401–410. https://doi.org/10.1037/h0056020.
Hudousková, Andrea, Eva Flanderková, Barbara Mertins & Kristýna Tomšů. 2014. Hypotéza vymazávání stop a její platnost v češtině u pacientů s Brocovou afázií. *Slovo a slovesnost* 19.
Hunt, Kellogg W. 1966. Recent Measures in Syntactic Development. *Elementary English*. National Council of Teachers of English 43(7). 732–739.
Ishkhanyan, Byurakn, Halima Sahraoui, Peter Harder, Jesper Mogensen & Kasper Boye. 2017. Grammatical and lexical pronoun dissociation in French speakers with agrammatic aphasia: A usage-based account and REF-based hypothesis. *Journal of Neurolinguistics* 44. 1–16. https://doi.org/10.1016/j.jneuroling.2017.02.001.
Jaecks, Petra, Martina Hielscher-Fastabend & Prisca Stenneken. 2012. Diagnosing residual aphasia using spontaneous speech analysis. *Aphasiology* 26(7). 953–970. https://doi.org/10.1080/02687038.2012.663075.
Jakobson, Roman. 1964. Towards a Linguistic Typology of Aphasic Impairments. In A. V .S. de Reuck & Maeve O'Connor (eds.), *Disorders of Language*, 21–46. J. & A. Churchill. https://doi.org/10.1002/9780470715321.ch3.
Jakobson, Roman. 1980. *Child Language, Aphasia and Phonological Universals*. (Trans.) Allan R. Keiler. 3rd edn. The Hague: Mouton Publishers.

Jakobson, Roman. 1995. Two Aspects of Language and Two Types of Aphasic Disturbances. In Linda R. Waugh & Monique Monville-Burston (eds.), *On Language*, 115–133. Harvard University Press.
Janda, Laura A. (ed.). 2013. *Cognitive linguistics: the quantitative turn: the essential reader* (Mouton Reader). Berlin: De Gruyter Mouton.
Janda, Laura A. 2019. Businessmen and Ballerinas Take Different Forms: A Strategic Resource for Acquiring Russian Vocabulary and Morphology. *Russian Language Journal* 69. 31–49.
Janda, Laura A. & Olga Lyashevskaya. 2011. Grammatical profiles and the interaction of the lexicon with aspect, tense, and mood in Russian. *Cognitive Linguistics* 22(4). https://doi.org/10.1515/cogl.2011.027.
Jayes, Mark & Rebecca Palmer. 2014. Initial evaluation of the Consent Support Tool: A structured procedure to facilitate the inclusion and engagement of people with aphasia in the informed consent process. *International Journal of Speech-Language Pathology* 16(2). 159–168. https://doi.org/10.3109/17549507.2013.795999.
Joanisse, Marc F. & James L. McClelland. 2015. Connectionist perspectives on language learning, representation and processing. *Wiley Interdisciplinary Reviews: Cognitive Science* 6(3). 235–247. https://doi.org/10.1002/wcs.1340.
Kagan, Aura & Mikael D. Z. Kimelman. 1995. Informed Consent in Aphasia Research: Myth or Reality? *Clinical Aphasiology* 23. 65–75.
Kalita, Z, M Zvolský, J Švancara & P Brabec. 2013. Srovnání epidemiologických dat u akutních cévních mozkových příhod podle metodiky ÚZIS a IKTA ve zlínském okrese a v ČR. *Česká a slovenská neurologie a neurochirurgie* 76/109(3). 350–357.
Kapatsinski, Vsevolod. 2010. Frequency of Use Leads to Automaticity of Production: Evidence from Repair in Conversation. *Language and Speech* 53(1). 71–105. https://doi.org/10.1177/0023830909351220.
Kapatsinski, Vsevolod. 2012. What statistics do learners track? Rules, constraints and schemas in (artificial) grammar learning. In Stefan Th. Gries & Dagmar Divjak (eds.), *Frequency Effects in Language Learning and Processing*. Berlin, Boston: DE GRUYTER. https://doi.org/10.1515/9783110274059.53.
Kapatsinski, Vsevolod & Joshua Radicke. 2009. *Frequency and the emergence of prefabs: Evidence from monitoring* (Typological Studies in Language). (Ed.) Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali & Kathleen Wheatley. *Formulaic Language: Volume 2. Acquisition, loss, psychological reality, and functional explanations* (Typological Studies in Language). John Benjamins Publishing Company. https://doi.org/10.1075/tsl.83.14kap.
Kemper, Susan, Donna Kynette, Shannon Rash, Kevin O'Brien & Richard Sprott. 1989. Life-span changes to adults' language: Effects of memory and genre. *Applied Psycholinguistics* 10(1). 49–66. https://doi.org/10.1017/S0142716400008419.
Kevická, Viktória, Jana Marková & Mária Gavalerová. 2020. Naratívny diskurz pacientov s Brocovou a anomickou afáziou. *Listy klinické logopedie*. Listy klinické logopedie 4(2). 113–123. https://doi.org/10.36833/lkl.2020.022.
Khudyakova, Mariya, Mira Bergelson, Yulia Akinina, Ekaterina Iskra, Svetlana Toldova & Olga Dragoy. 2016. Russian CliPS: a Corpus of Narratives by Brain-Damaged Individuals. In *Proceedings of LREC 2016 Workshop. Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016), Monday 23rd of May 2016*, 22–26.
Kim, Hana, Stephen Kintz, Kristen Zelnosky & Heather Harris Wright. 2019. Measuring word retrieval in narrative discourse: core lexicon in aphasia: Measuring word retrieval in narrative discourse. *International Journal of Language & Communication Disorders* 54(1). 62–78. https://doi.org/10.1111/1460-6984.12432.

Kiran, Swathi, David Caplan, Chaleece Sandberg, Joshua Levy, Alex Berardino, Elsa Ascenso, Sarah Villard & Yorghos Tripodis. 2012. Development of a Theoretically Based Treatment for Sentence Comprehension Deficits in Individuals With Aphasia. *American Journal of Speech-Language Pathology* 21(2). https://doi.org/10.1044/1058-0360(2012/11-0106).

Kittredge, Audrey K., Gary S. Dell, Jay Verkuilen & Myrna F. Schwartz. 2008. Where is the effect of frequency in word production? Insights from aphasic picture-naming errors. *Cognitive Neuropsychology* 25(4). 463–492. https://doi.org/10.1080/02643290701674851.

Klavan, Jane. 2012. *Evidence in linguistics: corpus-linguistic and experimental methods for studying grammatical synonymy*. Tartu: University of Tartu Press.

Klavan, Jane. 2017. Pitting corpus-based classification models against each other: a case study for predicting constructional choice in written Estonian. *Corpus Linguistics and Linguistic Theory* 0(0). https://doi.org/10.1515/cllt-2016-0010.

Klavan, Jane & Dagmar Divjak. 2016. The cognitive plausibility of statistical classification models: Comparing textual and behavioral evidence. *Folia Linguistica* 50(2). https://doi.org/10.1515/flin-2016-0014.

Klavan, Jane, Maarja-Liisa Pilvik & Kristel Uiboaed. 2015. The Use of Multivariate Statistical Classification Models for Predicting Constructional Choice in Spoken, Non-Standard Varieties of Estonian2. *SKY Journal of Linguistics* 28. 187–224.

Klavan, Jane & Ann Veismann. 2017. Are corpus-based predictions mirrored in the preferential choices and ratings of native speakers? Predicting the alternation between the Estonian adessive case and the adposition *peal* ‘on.’ *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 8(2). 59. https://doi.org/10.12697/jeful.2017.8.2.03.

Kolk, Herman. 1995. A Time-Based Approach to Agrammatic Production. *Brain and Language* 50(3). 282–303. https://doi.org/10.1006/brln.1995.1049.

Kolk, Herman & Claus Heeschen. 1992. Agrammatism, paragrammatism and the management of language. *Language and Cognitive Processes* 7(2). 89–129. https://doi.org/10.1080/01690969208409381.

Košťálová, Milena. 2012. Afázie a možnosti jejího skríninkového stanovení pomocí Mississippi Aphasia Screening Test - české verze (MASTcz). *Neurologie pro praxi* 13(6). 314–316.

Košťálová, Milena, Eva Bártková, Kateřina Šajgalíková, Alice Dolenská, Ladislav Dušek & Josef Bednařík. 2008. A standardization study of the Czech version of the Mississippi Aphasia Screening Test (MASTcz) in stroke patients and control subjects. *Brain Injury*. Taylor & Francis 22(10). 793–801. https://doi.org/10.1080/02699050802372190.

Košťálová, Milena, B Poláková, M Ulreichová, P Šmíd, E Janoušová, M Kuhn, J Klenková & J Bednařík. 2015. Dotazník funkcionální komunikace (DFK): validace originálního českého testu. *Česká a slovenská neurologie a neurochirurgie* 78/111(2). 188–195.

Kraljević, Jelena Kuvač, Gordana Hržica & Karolina Lice. 2017. CroDA: A Croatian Discourse Corpus of Speakers with Aphasia. *Hrvatska revija za rehabilitacijska istraživanja* 53(2). 61–71.

Křen, Michal, Václav Cvrček, T. Čapka, Anna Čermáková, M. Hnátková, Lucie Chlumská, T. Jelínek, et al. 2015. SYN2015: reprezentativní korpus psané češtiny. Ústav Českého národního korpusu FF UK. http://www.korpus.cz.

Křen, Michal, Václav Cvrček, Jan Henyš, M. Hnátková, T. Jelínek, J. Kocek, D. Kováříková, et al. 2020. SYN2020: reprezentativní korpus psané češtiny. Ústav Českého národního korpusu FF UK. www.korpus.cz.

Kynette, Donna & Susan Kemper. 1986. Aging and the loss of grammatical forms: a cross-sectional study of language performance. *Language & Communication* 6(1–2). 65–72. https://doi.org/10.1016/0271-5309(86)90006-6.

Lakoff, George. 1990. The Invariance Hypothesis: is abstract reason based on image-schemas? De Gruyter Mouton 1(1). 39–74. https://doi.org/10.1515/cogl.1990.1.1.39.
Langacker, Ronald W. 1988. A Usage-Based Model. In Brygida Rudzka-Ostyn (ed.), *Topis in Cognitive Linguistics*, 127–161. Amsterdam; Philadelphia, PA: John Benjamins.
Lehečková, Helena. 1985. Využití lingvistiky při rehabilitaci afázií. *Československá otolaryngologie* 34. 116–120.
Lehečková, Helena. 1986. Agramatismus v afázii. *Slovo a slovesnost* 47(2). 138–147.
Lehečková, Helena. 1988. Linguistic theories and the interpretation of agrammatism. *Clinical Linguistics & Phonetics*. Taylor & Francis 2(4). 271–289. https://doi.org/10.3109/02699208808985260.
Lehečková, Helena. 2001. Manifestation of aphasic symptoms in Czech. *Journal of Neurolinguistics* 14(2–4). 179–208.
Lehečková, Helena. 2008. Imperfect Language: How come? and So what? In Jouko Lindstedt, Andrew Chesterman, Mikhail Kopotev, Aila Laamanen, Ahti Nikunlassi, Juhani Nuorluoto, Jyrki Papinniemi, Pekka Pesonen & Johanna Viimaranta (eds.), *С любовью к слову. Festschrift in Honour of Professor Arto Mustajoki on the Occasion of his 60th Birthday* (Slavica Helsingiensia 35), 123–133. Helsinki: Department of Slavonic and Baltic Languages and Literatures, University of Helsinki.
Lehečková, Helena. 2009. Afázie jako zdroj poznatků o fungování jazyka. *Slovo a slovesnost*. Ústav pro jazyk český Akademie věd České republiky 70(1). 23–35.
Lehečková, Helena. 2012. Assessment and Treatment of Aphasia in Czech. *Folia Phoniatrica et Logopaedica* 64(4). 165–168. https://doi.org/10.1159/000340013.
Lehečková, Helena. 2016. Afázie v lingvistice, lingvistika v afáziologii. *Časopis pro moderní filologii* 98(1). 7–22.
Lesser, Ruth & Lesley Milroy. 1993. *Linguistics and Aphasia: Psycholinguistic and Pragmatic Aspects of Intervention*. London: Routledge.
Levshina, Natalia. 2018. Probabilistic grammar and constructional predictability: Bayesian generalized additive models of help + (to) Infinitive in varieties of web-based English. *Glossa: a journal of general linguistics* 3(1). https://doi.org/10.5334/gjgl.294.
Lind, Marianne, Kristian Emil Kristoffersen, Inger Moen & Hanne Gram Simonsen. 2009. Semi-spontaneous oral text production: Measurements in clinical practice. *Clinical Linguistics & Phonetics* 23(12). 872–886. https://doi.org/10.3109/02699200903040051.
Linnik, Anastasia, Roelien Bastiaanse & Barbara Höhle. 2016. Discourse production in aphasia: a current review of theoretical and methodological challenges. *Aphasiology* 30(7). 765–800. https://doi.org/10.1080/02687038.2015.1113489.
Loban, Walter. 1966. *Language Ability. Grades Seven, Eight, and Nine*. U.S. Department of Health, Education, and Welfare.
Lopatková, Markéta, V Kettnerová, A Vernerová, E Bejček & Z Žabokrtský. 2020. VALLEX 4.0. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-3524. (15 August, 2022).
Lounsbury, F G. 1954. Pausal, juncture and hesitation phenomena. In Charles E. Osgood & T A Sebeok (eds.), *Psycholinguistics: A survey of theory and research problems*, 93–101. Waverly Press.
Lukatela, G., B. Gligorijević, A. Kostić & M. T. Turvey. 1980. Representation of inflected nouns in the internal lexicon. *Memory & Cognition* 8(5). 415–423. https://doi.org/10.3758/BF03211138.
Luria, A.R. 1964. Factors and Forms of Aphasia. In A. V .S. de Reuck & Maeve O'Connor (eds.), *Disorders of Language*, 143–167. J. & A. Churchill. https://doi.org/10.1002/9780470715321.ch9.

Macková, Lucie, Ondřej Macko & Zuzana Hanibalová. 2020. AfaSlovník. https://www.afaslovnik.cz/. (15 August, 2022).
Maclay, Howard & Charles E. Osgood. 1959. Hesitation Phenomena in Spontaneous English Speech. *Word* 15(1). 19–44. https://doi.org/10.1080/00437956.1959.11659682.
MacWhinney, Brian. 2000. *The Childes Project: Tools for Analyzing Talk*. 3rd edn. Mahwah, N.J: Lawrence Erlbaum Associates. https://www.routledge.com/The-Childes-Project-Tools-for-Analyzing-Talk-Volume-II-the-Database/MacWhinney/p/book/9781138003491. (24 August, 2022).
MacWhinney, Brian. 2015. Emergentism. In Ewa Dąbrowska & Dagmar Divjak (eds.), *Handbook of Cognitive Linguistics*, 689–706. De Gruyter Mouton.
MacWhinney, Brian, Davida Fromm, Margaret Forbes & Audrey Holland. 2011. AphasiaBank: Methods for studying discourse. *Aphasiology* 25(11). 1286–1307. https://doi.org/10.1080/02687038.2011.589893.
MacWhinney, Brian, Davida Fromm, Audrey Holland, Margaret Forbes & Heather Wright. 2010. Automated analysis of the Cinderella story. *Aphasiology* 24(6–8). 856–868. https://doi.org/10.1080/02687030903452632.
Marini, Andrea, Sara Andreetta, Silvana del Tin & Sergio Carlomagno. 2011. A multi-level approach to the analysis of narrative language in aphasia. *Aphasiology* 25(11). 1372–1392. https://doi.org/10.1080/02687038.2011.584690.
Marshall, Jane. 2010. Classification of aphasia: Are there benefits for practice? *Aphasiology*. Routledge 24(3). 408–412. https://doi.org/10/cj5755.
Martínez-Ferreiro, Silvia, Byurakn Ishkhanyan, Vicente Rosell-Clarí & Kasper Boye. 2018. Prepositions and pronouns in connected discourse of individuals with aphasia. *Clinical Linguistics & Phonetics* 1–21. https://doi.org/10.1080/02699206.2018.1551935.
McWhorter, John. 2002. What happened to English? *Diachronica*. John Benjamins 19(2). 217–272. https://doi.org/10.1075/dia.19.2.02wha.
Menn, Lise & Loraine K. Obler (eds.). 1990. *Agrammatic aphasia: A cross-language narrative sourcebook*. Amsterdam ; Philadelphia, PA: John Benjamins Publishing Company.
Messerschmidt, Maria, Kasper Boye, Mathilde Marie Overmark, Sofie Theilmann Kristensen & Peter Harder. 2018. Sondringen mellem grammatiske og leksikalske præpositioner. *Ny forskning i grammatik* (25). https://doi.org/10.7146/nfg.v26i25.109304.
Nakase-Thompson, R., E. Manning, M. Sherer, S. A. Yablon, S. L. T. Gontkovsky & C. Vickery. 2005. Brief assessment of severe language impairments: Initial validation of the Mississippi aphasia screening test. *Brain Injury*. Taylor & Francis 19(9). 685–691. https://doi.org/10.1080/02699050400025331.
Nozari, Nazbanou, Audrey K. Kittredge, Gary S. Dell & Myrna F. Schwartz. 2010. Naming and repetition in aphasia: Steps, routes, and frequency effects. *Journal of Memory and Language* 63(4). 541–559. https://doi.org/10/bbkggw.
Olness, Gloria Streit. 2006. Genre, verb, and coherence in picture-elicited discourse of adults with aphasia. *Aphasiology* 20(2–4). 175–187. https://doi.org/10.1080/02687030500472710.
Onnis, Luca & Falk Huettig. 2021. Can prediction and retrodiction explain whether frequent multi-word phrases are accessed 'precompiled' from memory or compositionally constructed on the fly? *Brain Research* 147674. https://doi.org/10/gmz7qn.
Palmer, Rebecca & Gail Paterson. 2011. One Size Does Not Fit All: Obtaining informed consent from people with aphasia. *Advances in Clinical Neuroscience and Rehabilitation* 11(2). 30–31.
Panevová, Jarmila. 1980. *Formy a funkce ve stavbě české věty*. Praha: Academia.
Paradis, Michel (ed.). 2001. *Manifestations of Aphasia Symptoms in Different Languages*. Emerald Group Publishing.

Pelucchi, Bruna, Jessica F. Hay & Jenny R. Saffran. 2009. Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition* 113(2). 244–247. https://doi.org/10.1016/j.cognition.2009.07.011.
Perek, Florent & Adele E. Goldberg. 2017. Linguistic generalization on the basis of function and constraints on the basis of statistical preemption. *Cognition* 168. 276–293. https://doi.org/10.1016/j.cognition.2017.06.019.
Perkins, Lisa, Jenni Crisp & David Walshaw. 1999. Exploring conversation analysis as an assessment tool for aphasia: the issue of reliability. *Aphasiology* 13(4–5). 259–281. https://doi.org/10.1080/026870399402091.
Pick, Arnold. 1913. *Die Agrammatischen Sprachstörungen: Studien zur Psychologischen Grundlegung der Aphasielehre* (Monographien aus dem Gesamtgebiete der Neurologie und Psychiatrie). Vol. 7. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-26371-6.
Pickering, Martin J. & Victor S. Ferreira. 2008. Structural priming: A critical review. *Psychological Bulletin* 134(3). 427–459. https://doi.org/10.1037/0033-2909.134.3.427.
Presson, Nora & Brian MacWhinney. 2011. The Competition Model and Language Disorders. In *The Handbook of Psycholinguistic and Cognitive Processes*. Routledge. https://doi.org/10.4324/9780203848005.ch2.
Prins, Ronald & Roelien Bastiaanse. 2004. Analysing the spontaneous speech of aphasic speakers. Review. *Aphasiology* 18(12). 1075–1091. https://doi.org/10.1080/02687030444000534.
Pritchard, Madeleine, Katerina Hilari, Naomi Cocks & Lucy Dipper. 2017. Reviewing the quality of discourse information measures in aphasia: Reviewing the quality of discourse information measures in aphasia. *International Journal of Language & Communication Disorders* 52(6). 689–732. https://doi.org/10/gfkm6g.
Pritchard, Madeleine, Katerina Hilari, Naomi Cocks & Lucy Dipper. 2018. Psychometric properties of discourse measures in aphasia: acceptability, reliability, and validity: Psychometric properties of discourse measures in aphasia. *International Journal of Language & Communication Disorders* 53(6). 1078–1093. https://doi.org/10.1111/1460-6984.12420.
Ramscar, Michael & Robert Port. 2015. Categorization (without categories). In Ewa Dabrowska & Dagmar Divjak (eds.), *Handbook of Cognitive Linguistics*, 75–99. De Gruyter Mouton.
Rochon, Elizabeth, Eleanor M. Saffran, Rita Sloan Berndt & Myrna F. Schwartz. 2000. Quantitative Analysis of Aphasic Sentence Production: Further Development and New Data. *Brain and Language* 72(3). 193–218. https://doi.org/10.1006/brln.1999.2285.
Roo, Esterella de, Herman Kolk & Ben Hofstede. 2003. Structural properties of syntactically reduced speech: A comparison of normal speakers and Broca's aphasics. *Brain and Language* 86(1). 99–115. https://doi.org/10.1016/S0093-934X(02)00538-2.
Rosch, Eleanor H. 1973. On the Internal Structure of Perceptual and Semantic Categories. In Thomas E. Moore (ed.), *Cognitive Development and Acquisition of Language*, 111–144. New York: Academic Press. https://doi.org/10.1016/B978-0-12-505850-6.50010-4.
Rossi, Eleonora & Roelien Bastiaanse. 2008. Spontaneous speech in Italian agrammatic aphasia: A focus on verb production. *Aphasiology* 22(4). 347–362. https://doi.org/10.1080/02687030701407093.
Saffran, Eleanor M., Rita Sloan Berndt & Myrna F. Schwartz. 1989. The quantitative analysis of agrammatic production: Procedure and data. *Brain and Language* 37(3). 440–479. https://doi.org/10.1016/0093-934X(89)90030-8.
Schmid, Hans-Jörg (ed.). 2016. *Entrenchment and the psychology of language learning: how we reorganize and adapt linguistic knowledge* (Language and the Human Lifespan). First Edition. Washington, DC: American Psychological Association.
Schmid, Hans-Jörg & Helmut Küchenhoff. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and

cognitive underpinnings. *Cognitive Linguistics* 24(3). https://doi.org/10.1515/cog-2013-0018.
Schneider, Ulrike. 2016. Hesitation placement as evidence for chunking: A corpus-based study of spoken English. In *Experience Counts: Frequency Effects in Language*, 61–90. De Gruyter. https://doi.org/10.1515/9783110346916-004.
Šebesta, Karel, Z. Bedřichová, K. Šormová, B. Štindlová, M. Hrdlička, M. Hrdličková, V. Petkevič, et al. 2014. CzeSL-SGT: korpus češtiny nerodilých mluvčích s automaticky provedenou anotací, verze 2 z 28. 7. 2014. Ústav Českého národního korpusu FF UK. http://www.korpus.cz. (24 August, 2022).
Sedova, Petra, Robert D. Brown, Miroslav Zvolsky, Pavla Kadlecova, Tomas Bryndziar, Tomáš Kubelka, Viktor Weiss, Ondřej Volný, Josef Bednarik & Robert Mikulik. 2017. Incidence of Hospitalized Stroke in the Czech Republic: The National Registry of Hospitalized Patients. *Journal of Stroke and Cerebrovascular Diseases* 26(5). 979–986. https://doi.org/10.1016/j.jstrokecerebrovasdis.2016.11.006.
Slobin, Dan I. 1991. Aphasia in Turkish: Speech production in Broca's and Wernicke's patients. *Brain and Language* 41(2). 149–164. https://doi.org/10.1016/0093-934X(91)90150-Y.
Smolash, Ruth Rachel. 1997. *Conversational turn loss due to silent pauses in aphasic speech*. Boston, MA: Northeastern University MA Thesis.
Sophie, Bustamante, Claire Croteau & Justine Tremblay-Beausejour. 2007. Conversational pauses between persons with aphasia and their spouses.
Springer, Luise, Walter Huber, Klaus-J. Schlenck & Claudia Schlenck. 2000. Agrammatism: Deficit or compensation? Consequences for aphasia therapy. *Neuropsychological Rehabilitation* 10(3). 279–309. https://doi.org/10.1080/096020100389165.
Straková, Jana, Milan Straka & Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 13–18. Baltimore, Maryland: Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-5003.
Street, James A. & Ewa Dąbrowska. 2010. More individual differences in language attainment: How much do adult native speakers of English know about passives and quantifiers? *Lingua* 120(8). 2080–2094. https://doi.org/10.1016/j.lingua.2010.01.004.
Taylor, John R. 1995. *Linguistic categorization: prototypes in linguistic theory*. 2nd ed. London : New York: Clarendon Press ; Oxford University Press.
Taylor, John R. 2012. *The mental corpus: how language is represented in the mind*. Oxford ; New York: Oxford University Press.
Tesak, Juergen & Chris Code. *Milestones in the History of Aphasia: Theories and Protagonists*. https://www.routledge.com/Milestones-in-the-History-of-Aphasia-Theories-and-Protagonists/Tesak-Code/p/book/9780415655576. (16 August, 2022).
Těšitelová, Marie. 1985. *Kvantitativní charakteristiky současné češtiny*. Praha: Academia.
Thompson, Cynthia K. 2003. Unaccusative verb production in agrammatic aphasia: the argument structure complexity hypothesis. *Journal of Neurolinguistics* 16(2–3). 151–167. https://doi.org/10.1016/S0911-6044(02)00014-3.
Thompson, Cynthia K. & Jungwon Janet Choy. 2009. Pronominal Resolution and Gap Filling in Agrammatic Aphasia: Evidence from Eye Movements. *Journal of Psycholinguistic Research* 38(3). 255–283. https://doi.org/10.1007/s10936-009-9105-7.
Thompson, Cynthia & Lewis Shapiro. 2005. Treating agrammatic aphasia within a linguistic framework: Treatment of Underlying Forms. *Aphasiology* 19(10–11). 1021–1036. https://doi.org/10.1080/02687030544000227.
Thompson, Susan P & Elissa L Newport. 2007. Statistical Learning of Syntax: The Role of Transitional Probability. *Language Learning and Development* 3(1). 1–42.

Tummers, Jose, Kris Heylen & Dirk Geeraerts. 2005. Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory*. De Gruyter Mouton 1(2). 225–261. https://doi.org/10/b5fbj5.
Ulatowska, Hanna K., Lee Allard & Sandra Bond Chapman. 1990. Narrative and Procedural Discourse in Aphasia. In Yves Joanette & Hiram H. Brownell (eds.), *Discourse Ability and Brain Damage*, 180–198. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4612-3262-9_8.
Valdois, Sylviane, Johns Ryalls & AndréRoch Lecours. 1989. Luria's aphasiology: A critical review. *Journal of Neurolinguistics* 4(1). 37–63. https://doi.org/10/cp87rd.
Varlokosta, Spyridoula, Spyridoula Stamouli, Athanasios Karasimos, Maria Kakavoulia, Michaela Nerantzini, Aikaterini Pantoula, Alexandra Economou & Athanassios Protopapas. 2016. A Greek Corpus of Aphasic Discourse: Collection, Transcription, and Annotation Specifications. 8.
Verhagen, Véronique. 2020. *Illuminating variation: Individual differences in entrenchment of multi-word units*. Tilburg: Tilburg University Dissertation. https://www.lotpublications.nl/illuminating-variation-individual-differences-in-entrenchment-of-multi-word-units. (11 February, 2020).
Verhagen, Véronique, Maria Mos, Ad Backus & Joost Schilperoord. 2018. Predictive language processing revealing usage-based variation. *Language and Cognition*. Cambridge University Press 10(2). 329–373. https://doi.org/10/ggnz72.
Wackernagel, Jacob, George Walkden, Christina Sevdali & Morgan Macleod. 2020. *On a law of Indo-European word order*. *Language Science Press*. Language Science Press. https://doi.org/10.5281/zenodo.3978908.
Wagenaar, Erin, Catherine Snow & Ronald Prins. 1975. Spontaneous speech of aphasic patients: A psycholinguistic analysis. *Brain and Language* 2. 281–303. https://doi.org/10.1016/S0093-934X(75)80071-X.
Walker, Virginia G., Patricia M. Roberts & Dona Lea Hedrick. 1988. Linguistic Analyses of the Discourse Narratives of Young and Aged Women. *Folia Phoniatrica et Logopaedica* 40(2). 58–64. https://doi.org/10.1159/000265885.
Warren, Tessa, Michael Walsh Dickey & Teljer L. Liburd. 2017. A rational inference approach to group and individual-level sentence comprehension performance in aphasia. *Cortex* 92. 19–31. https://doi.org/10.1016/j.cortex.2017.02.015.
Wernicke, Carl. 1974. *Der aphasische Symptomenkomplex*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-65950-8.
Westerhout, E N. 2005. *A Corpus of Dutch Aphasic Speech: Sketching the Design and Performing a Pilot Study*. Utrecht University MA Thesis. http://dspace.library.uu.nl/bitstream/handle/1874/8357/Thesis.pdf?sequence=1&isAllowed=y. (17 August, 2021).
Westerhout, Eline & Paola Monachesi. 2007. A pilot study for a Corpus of Dutch Aphasic Speech (CoDAS). In Khalil Sima'an, Maarten de Rijke, Remko Scha & Rob van Son (eds.), *Proceedings of the Sixteenth Computational Linguistics in the Netherlands*, 83–98.
Wilkinson, Ray, Karen Bryan, Sarah Lock, Kate Bayley, Jane Maxim, Carolyn Bruce, Anne Edmundson & Diana Moir. 1998. Therapy Using Conversation Analysis: Helping Couples adapt to Aphasia in Conversation. *International Journal of Language & Communication Disorders* 33(S1). 144–149. https://doi.org/10.3109/13682829809179412.
Willems, Klaas. 2012. Intuition, introspection and observation in linguistic inquiry. *Language Sciences* (Converging Data Sources in Cognitive Linguistics) 34(6). 665–681. https://doi.org/10.1016/j.langsci.2012.04.008.
Williams, Caroline, Andrew Thwaites, Paula Buttery, Jeroen Geertzen, Billi Randall, Meredith Shafto, Barry Devereux & Lorraine Tyler. 2010. The Cambridge Cookie-Theft Corpus: A

Corpus of Directed and Spontaneous Speech of Brain-Damaged Patients and Healthy Individuals. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/327_Paper.pdf. (24 August, 2022).

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann & Han Sloetjes. 2006. ELAN: a Professional Framework for Multimodality Research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf. (24 August, 2022).

Wulfeck, Beverly, Elizabeth Bates & Rita Capasso. 1991. A crosslinguistic study of grammaticality judgments in Broca's aphasia. *Brain and Language* 41(2). 311–336. https://doi.org/10.1016/0093-934X(91)90158-W.

Zeschel, Arne. 2008. Lexical chunking effects in syntactic processing. *Cognitive Linguistics* 19(3). https://doi.org/10.1515/COGL.2008.016.

Zíková, Magdalena. 2018. Anaforická funkce výrazu ten v gramatikalizační perspektivě. *Studie z aplikované lingvistiky / Studies in Applied Linguistics* 9(Zvláštní číslo). 106–213.

Zimmerer, V. C., L. Newman, R. Thomson, M. Coleman & R. A. Varley. 2018. Automated analysis of language production in aphasia and right-hemisphere damage: frequency and collocation strength. *Aphasiology* 32(11). 1267–1283. https://doi.org/10.1080/02687038.2018.1497138.

2015. Time and again: The changing effect of word and multiword frequency on phonetic duration for highly frequent sequences. *The Mental Lexicon* 9(3). 377–400. https://doi.org/10.1075/ml.9.3.01arn.

2017. ORAL: korpus neformální mluvené češtiny. Prague: Ústav Českého národního korpusu FF UK. www.korpus.cz.

7 Appendix 2: Informed consent form

7.1.1 Informovaný souhlas se zařazením do projektu “Příprava korpusu afatické řeči”

Michal Láznička

Ústav obecné lingvistiky

Filozofická fakulta Univerzity Karlovy

Náměstí Jana Palacha xy, 11000 Praha

michal.laznicka@ff.cuni.cz

731414997

Kdo jsme?

- Badatelský tým studentů obecné jazykovědy na FF UK: M. Láznička, V. Diatka, T. Papayová, K. Vaníčková
- Nás projekt podporuje program vnitřních grantů FF UK

O co nám jde?

- V rámci projektu pořizujeme nahrávky osob s afázií.
- Vytvoříme elektronickou databázi textů.
- Nahrávky zpracujeme a vložíme do databáze.
- Databázi zpřístupníme přes webové rozhraní dalším specialistům

Proč to děláme?

- Chceme zlepšit naše poznání češtiny v afázii
- Chceme toto poznání zpřístupnit odborníkům a studentům
- Chceme pomoci ve výuce a terapii
- Chceme prohloubit naše chápání jazyka obecně

Jak nám můžete pomoci?

- Účastí ve výzkumu přispějete ke vzniku databáze
- Vaše účast je zcela dobrovolná
- Pokud se nezúčastníte, nehrozí žádný postih

Co bude vašim úkolem?

- Zúčastníte se dvou rozhovorů
- Každý rozhovor bude trvat asi 30 minut
- K rozhovorům se setkáme u vašeho terapeuta
- Rozhovory proběhnou v různé dny
- Rozhovory budu nahrávat videokamerou
- Během našeho rozhovoru se vás budu ptát na různé věci
- Na mé otázky nemusíte odpovědět, když nebudete chtít
- Pustím vám také krátké video a ukážu vám několik obrázků
- Dám vám otázky k videu a k obrázkům
- Vaše odpovědi nebudu nijak hodnotit

Co se bude dít s nahrávkou dál?

- Nahrané rozhovory přepíšeme do textových souborů
- Přepisy a zvukové nahrávky umístíme do databáze
- Databáze bude přístupná pouze registrovaným uživatelům
- Přístup umožníme pouze odborníkům a studentům
- Nahrávky budou k dispozici pro účely výzkumu a výuky
- Do přepisů nezahrneme vaše jméno, ani jména vašich blízkých osob
- Přepisy však budou obsahovat některé vaše osobní údaje:
    - věk
    - pohlaví
    - region, z kterého pocházíte
    - nejvyšší dosažené vzdělání
    - informace o vašich obtížích
- V databázi bude slyšet váš hlas

Nechci, aby byl můj záznam v databázi, můžu se přesto zúčastnit?

- Ano, i v takovém případě můžete pomoct
- Rozhovor bude probíhat stejně
- Rozhovor budu nahrávat na kameru
- Rozhovor nepřepíšeme a nezařadíme do databáze
- Nahrávku si ponechám pro účely svého vlastního výzkumu
- Pokud nahrávku využiju, budu z ní citovat pouze krátké úryvky textu
- V ukázkách nebude vaše jméno, ani jména vašich blízkých osob
- Zvuk ani obraz nebude nikde přístupný
- Pokud nahrávku použiju, budu citovat některé vaše osobní údaje:
    - věk
    - pohlaví
    - informace o vašich obtížích

Jsou nějaká rizika?

- Účast ve výzkumu vás nijak neohrozí
- Pokud se unavíte, rozhovor přerušíme
- Pokud vám budou otázky nepříjemné, nemusíte odpovídat, nebo můžeme rozhovor ukončit

Čemu to prospěje?

- Pomůžete ve výzkumu
- Pomůžete zlepšit naše chápání jazykových poruch
- Pomůžete nám pochopit, jak funguje čeština
- Nejde o terapii, účast ve výzkumu nepřispěje k dřívějšímu zlepšení

Co když si to rozmyslím?

- Pokud si svou účast rozmyslíte, rozhovor můžete kdykoli ukončit
- Pokud rozhovor ukončíte, nehrozí vám žádný postih
- Pokud budete chtít vyjmout svou nahrávku z databáze, kontaktujte mě
- Pokud budete chtít svou nahrávku odstranit, nehrozí vám žádný postih

Co když budu chtít o výzkumu zjistit více?

- Pokud se chcete dozvědět více, kontaktujte mě

Souhlasím se zařazením nahrávky do databáze?

- ano
- ne

Já, ……., potvrzuji svým podpisem, že rozumím výše uvedeným informacím a souhlasím se svou účastí v tomto výzkumu.

V …….. dne…….

…….

podpis