THE UNIVERSITY *of* EDINBURGH

Automatic Detection of Clusters and Switches in Turkish Semantic Verbal Fluency Data

Rabia Yasa Kostas

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2024

Abstract

Verbal fluency tests are popular measures of executive function. These tests involve listing as many words from a given category as possible in a short time, typically 60 seconds. In phonemic verbal fluency tests, these words should begin with the same letter; in semantic verbal fluency tests (SVF), they should belong to the same category, e.g., animals. SVF is quick to administer, amenable to semi-automated analysis, and can be used to screen for cognitive impairments such as dementia. Troyer and collaborators proposed a fine-grained analysis method for SVF sequences that divides them into clusters, i.e., sequences of more closely semantically related words. Useful metrics that can be derived from such an analysis include mean cluster size and the number of switches between clusters. The aim of this thesis is to develop semi-automated methods to extract cluster- and switch-related metrics from Turkish SVF sequences.

First, we conducted a systematic review of studies that report SVF performance of healthy adult native Turkish speakers, using international and Turkish databases including unpublished theses. We particularly focused on normative data and commonly used methods for collecting and analysing SVF data. We found that all included papers reported SVF sequences using the animal category, followed by first names. Considering the size of Turkish diaspora, there was a lack of studies comparing monolingual speakers to bilingual speakers. Detailed analyses beyond word count, such as perseverations, category violations, and clustering/switching were only rarely reported. Semi-automatic and automatic approaches were almost never used. The thesis therefore fills a clear gap in the literature.

For our work on Turkish, we chose two computational approaches that can be easily adapted to languages with comparatively few corpus resources: a simple bigram method and a vector-space model (word2vec). We initially implemented and tested those methods on a Spanish dataset which included 50 healthy participants and 14 participants diagnosed with familial AD. Both computational models positioned switches very similarly to manual annotations, achieving $F1$=0.756 for Bigram and $F1$=0.8309 for Word2vec. There is no difference in terms of cluster sizes ($p>0.01$), but healthy participants produce significantly more switches ($p<0.001$). These findings hold both for the manual analysis and the automatic analysis.

Since there are no public datasets of Turkish SVF data, we collected SVF data online from native speakers of Turkish with no self-reported cognitive impairments living both in Turkey and abroad. To the best of our knowledge, this is the first online

spoken corpus of SVF for Turkish. The study used the three most frequently used categories in Turkish SVF data that have also been reported for other languages, namely animals, fruits and vegetables, and supermarket items. The study had two parts, an initial Qualtrics survey for screening and collecting relevant participant information, and a web-based app for collecting three SVF sequences. 286 participants consented to take part in the survey, and 137 (47.9%) continued on to the SVF app. In total, we collected 311 SVF sequences (Animals=105, Vegetables and Fruits=105, Supermarket Items=101) from 137 adults. The mean number of items produced per category is 25.04 (SD=X) for animals, 25.32 (SD=Y) for fruits and vegetables, and 25.97 (SD=Z) for supermarket items. Overall, data quality of the recorded sequences was good. The reasons for the drop off between survey and SVF data collections need to be investigated in further work.

Finally, we adapted the computational techniques used for Spanish to the Turkish SVF data and assessed their ability to replicate clustering and switching based metrics. We found that both bigram and word2vec performed satisfactorily. There was no significant difference in cluster sizes, and switch numbers were highly correlated ($p<0.001$). In terms of predicting switch position, word2vec reached $F1=0.738$ and Bigram achieved $F1=0.66$. Next, we examined whether findings obtained from manual annotation of clusters and switches could be replicated using metrics derived using the two computational methods. Specifically, we investigated cluster size and switch numbers between male and female participants (sex) and between mono- and multilingual participants (multilinguality). Based on the manual analysis, we established that male participants created larger clusters than female participants, but used a similar number of switches. There were no significant differences between monolingual and multilingual participants. Both findings are in line with the existing literature on Turkish SVF. While bigram and word2vec yielded a similar result regarding number of switches, only word2vec-derived metrics replicated the difference in cluster size between male and female participants.

In future work, other computational approaches, such as large language models, should be explored, automatic speech recognition should be integrated to eliminate the need for manual transcription, and additional speech-based features can be investigated. Finally, user experience research may help to improve online data collection and reduce the number of participants who drop out of the study before speech data collection.

Lay summary

Semantic verbal fluency (SVF) tests give individuals a limited amount of time to list as many words as they can within a specific category, like animals. These tests help us understand how people recall words from their memory. Our research revealed that studies on SVF in native Turkish speakers often rely on traditional manual methods like word counting to evaluate SVF tests. Consequently, our goal was to fill this gap in Turkish research by adapting computer-based automatic evaluation methods commonly employed in other languages. To achieve this, we initially applied automated methods to Spanish, a widely studied language. We then collected our own data for the Turkish language and introduced a comprehensive and reproducible analysis approach. We developed a data collection application that is accessible via phone or computer, which makes data collection more convenient for both participants and researchers and allowed us to gather data from both within and outside Türkiye. We intended for our data collection method and analysis techniques to serve as a pioneering resource for future Turkish language studies in the field of SVF.

Acknowledgements

I would like to express my heartfelt gratitude to my supervisors, Maria and Sarah, for their pivotal roles in guiding me through the PhD journey at the University of Edinburgh. Maria, I'm truly thankful for leading the way and for always being there when I needed an assistance. I am grateful for your huge support, not only in shaping my academic journey, but also in providing me with moral and encouragement in times of illness and health. I am incredibly fortunate to have you as my principal supervisor, and words cannot adequately express my gratitude. Sarah, thank you for introducing me to the fascinating world of cognitive science and guiding me to discover about topics I was initially unfamiliar with. Thank you for the incredible support provided especially during the challenging times of the Covid-19 pandemic. Both of your dedication to minimizing impact of pandemic on my research have been make this thesis possible.

I want to express my deepest thanks to my family for always believing in me. Mom and dad, your love, encouragement, and support in me have sustained me through the challenges of this journey. Your unwavering faith in my abilities has been a constant source of strength and motivation. I will forever have a deep gratitude and respect for the sacrifices you made to enable me to pursue my education. Sister and Brother, you have always inspired me with your achievements, the respect and love you have shown me have always made me feel stronger than I am. I am deeply grateful for the role you have played in my personal development.

I would like to thank to all the friends I have met in the lab, office, and social circles at the university and in Scotland. You are the ones who have motivated and cheered me throughout this story, while being far away from my homeland. My sincere thanks goes out to each one of you individually.

Lastly but most importantly, I would like to express my heartfelt gratitude to my dear husband. You are my hero. I am so lucky to have you and thankful for your unwavering presence in every moment of my life. For nearly two decades, we have stood side by side, growing and learning together. Our belief has made this journey possible, just as it has in the past, and we will continue to create lovely memories together.

Declaration

I declare that this thesis was composed by myself, and that this work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work that has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated below.

Chapter 4 was submitted to the Journal of Clinical and Experimental Neuropsychology (CEN-RA 23-107) by Rabia Yasa Kostas, Kahraman Kostas, Sarah E. MacPherson, and Maria K. Wolters. Study received the first decision and invited to re-submit. The chapter contains the revised text. The preliminary findings were presented as an abstract in Psychonomic Society: 61st Annual Virtual Meeting, November 19-22, 2020.

In Chapter 5 original Troyer animal taxonomy was adapted into Spanish language with the help of bilingual Spanish and English speaker: Charlotte Sudduth.

In Chapter 6, web-based audio recording application was co-designed by Rabia Yasa Kostas, Sarah E. MacPherson, and Maria K. Wolters. The English version of an application was implemented by Danyi He, as a part of her MSc degree requirement of University of Edinburgh, under the supervision of Maria K. Wolters and tutoring by Rabia Yasa Kostas. The application was translated into Turkish version and conducted on the Google Cloud Platform by Rabia Yasa Kostas. Abstract of the study was accepted for 64th Annual Meeting of Psychonomic Society and will be presented in November 16-19, 2023 at San Francisco, California, USA.

This thesis was supported by Republic of Türkiye Ministry of National Education.

(*Rabia Yasa Kostas*)

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation		Definition
EF(s)	:	Executive functions
D-KEFS	:	Delis-Kaplan Executive Function System
CERAD	:	Consortium to Establish a Registry for Alzheimer's Disease
VFT	:	Verbal fluency test
SVF	:	Semantic verbal fluency
PVF	:	Phonemic verbal fluency
COWAT	:	Controlled Oral Word Association Test
AD	:	Alzheimer Disease
PSEN-1	:	Presenilin-1 gene
ASR	:	Automatic speech recognition
MCAS	:	The Minnesota Cognitive Acuity Screen
MoCA	:	Montreal Cognitive Assessment
CLSA	:	The Canadian Longitudinal Study on Aging
BHR	:	The Brain Health Registry
TUIK	:	Turkish Statistical Institute
MCI	:	Mild cognitive impairment
LEAP-Q	:	The Language Experience and Proficiency Questionnaire
LSBQ	:	The Language and Social Background Questionnaire
POS	:	Part of speech
BOW	:	Bag of Words
Tf	:	Term Frequency
LSA	:	Latent Semantic Analysis
TF-IDF	:	Term Frequency-Inverse Document Frequency
PMI	:	Pointwise Mutual Information
CBOW	:	Continuous Bag-of-Words (Word2Vec)
MMSE	:	Mini Mental State Examination
ASD	:	Autism Spectrum Disorder
AMCI	:	Amnestic Mild Cognitive Impairment
LSTM	:	Long Short-Term Memory
RNN	:	Recurrent Neural Network
GPT	:	Generative Pre-Training by OpenAI Transformer
BERT	:	Bidirectional Encoder Representations from Transformers
SVM	:	Support Vector Machine
Prospero	:	International Prospective Register of Systematic Reviews
PI(E)COS	:	The Cochrane framework stands for Population, Intervention, Comparison, Outcome and Study.

Chapter 1

Introduction

.

1.1 What is Semantic Verbal Fluency?

In neuropsychological assessment, validated collections of tests with highly standardised stimuli and tasks are used to gain insight into a person's neurocognitive function.

Verbal fluency tests (VFTs) are assessment tools widely used to measure executive function (Henry and Crawford, 2005). Executive function (EF) includes the set of cognitive processes that are important for performing goal-directed behaviours. These include managing attention in daily life, problem solving in current situations, planning for the future, and switching between activities Gilbert and Burgess (2008). VFTs are part of standardised cognitive assessment batteries such as Addenbrooke's Cognitive Examination (Mathuranath et al., 2000a), the Delis-Kaplan Executive Function System (D-KEFS) (Delis et al., 2001b), and the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) (Morris et al., 1989).

A VFT requires participants to produce as many words that fulfil a given condition as they can in a short amount of time. In addition to executive function, performance on verbal fluency tests is also affected by other aspects such as the organisation of long term memory (Mathuranath et al., 2003). There are two main variations of VFTs (Lezak et al., 2004):

Semantic verbal fluency (SVF) tests , also called category fluency tests, require participants to produce words belonging to a particular semantic category, such as animals, fruits and vegetables, household items etc.

Phonemic verbal fluency (PVF) tests , also called letter fluency tests or Controlled Oral Word Association Test (COWAT), require participants to produce words starting with a given letter—typically F, A, or S in English (Beatty, 2002), but the letters can change depending on the language. For example, in Turkish, either A, E, and Z or K, A, and S are used (Sumiyoshi et al., 2014a).

Additionally, there is a mixed type of fluency, **alternate fluency**. Tests of alternate fluency ask participants to switch between phonemic and semantic fluency tasks. For example, participants might be given the instructions 'animals then vegetables', 'words beginning with the letter F then A', or 'words beginning with the letter S, then supermarket items' (Villalobos et al., 2022).

While poor SVF performance is related to damage of the temporal lobes of the brain, decreased PVF performance is related to frontal lobe lesions (Lopes et al., 2009b; Hazin et al., 2016). Figure 1.1 illustrates the main regions of the brain and their function.

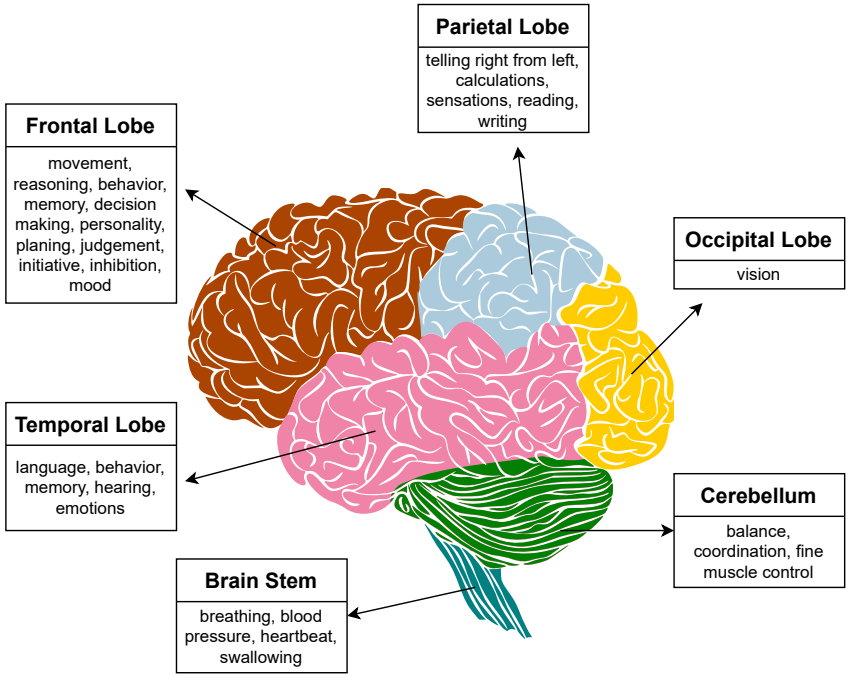


Figure 1.1: Brain regions and functions. Figure was adapted from the webpage of the National Brain Tumor Society (2023).

Since VFTs are quick and easy to administer, they are excellent candidates for neuropsychological screening. For example, poor SVF performance compared to healthy controls has been found in people with Alzheimer's disease, Parkinson's disease, and Huntington's disease (Randolph et al., 1993); mild cognitive impairment (Paula et al.,

2018); traumatic brain injury (Woods et al., 2016a); temporal lobe lesions (Troyer et al., 1998a); and HIV infection and chronic alcoholism (Fama et al., 2011).

The most straightforward and common approach to evaluate SVF performance is counting the number of words generated within a specified amount of time. Two additional metrics are also widely used: perseverations (repeated words) and category violations (words produced that do not fit into the desired category). However, additional metrics can be derived, for example by taking into account the semantic relationships between the words produced. Troyer et al. (1997) proposed a now widely used method for determining clusters of semantically related words and defined metrics based on those clusters, such as the number of switches between clusters and the mean cluster size (c.f. Section 2.4.2). Troyer also provided a comprehensive animal taxonomy for analysing sequences of words produced for the animal SVF task.

There are a few drawbacks when applying the Troyer method in clinical practice. Manual annotation of clusters and switches is time consuming, and the official Troyer taxonomy does not cover all types of animals. It also misses culture-specific groups of animals Kim et al. (2019).

Computational linguistic methods have also been used to uncover the internal structure of SVF sequences and derive clinically relevant metrics. While some approaches have focused on automating manual clustering and switching analyses (e.g. Kim et al. (2019)), others have proposed new metrics and validated them on standard corpora of SVF data (e.g. Pakhomov et al. (2012); Linz et al. (2017b)).

1.2 Semantic Verbal Fluency in Turkish

The semantic verbal fluency test has been adapted into many different languages. Normative studies exist in English (Tombaugh et al., 1999), Spanish (Benito-Cuadrado et al., 2002b), Arabic (Khalil, 2010a), Chinese (Feng et al., 2012), Dutch (Van Der Elst et al., 2006b), Hebrew (Kavé, 2005), Greek (Kosmidis et al., 2004b), Persian (Ghasemian-shirvan et al., 2018), and many more.

Turkish refers to the language spoken in Türkiye (Turkey[1]), which is a member of the Turkic language family and accounts for 30% of all speakers of Turkic lan-

[1]After an official letter from the Turkish government requested the use of 'Türkiye' rather than 'Turkey', the United Nations agreed to change the country's name and declared the new spelling in June 2022. For this reason, Türkiye will be used as the name of the country in this study. The official announcement of the United Nations Türkiye can be found here: https://turkiye.un.org/en/184798-turkeys-name-changed-t%C3%BCrkiye

guages (Kornfilt, 1990). Turkic languages are spoken by about 200 million people all over the world, more than 80 percent of whom speak the following languages: Turkish, Azerbaijani, Uzbek, Kazakh, and Uyghur (Rybatzki, 2020). Turkish is the official language in Türkiye, which has a population of 85 million based on a report of Turkish Statistical Institute (TUIK, 2022), and the Northern part of Cyprus, where it is spoken by 286,257 people according to the last census in 2011 (TRNC, 2011). The Turkish diaspora comprises 6.5 million people around the world (Turkish Ministry of Foreign Affairs, 2022), many of whom live in the Balkans and Greece (Kornfilt, 1990; Ak, 2023). Given that almost 100 million people speak Turkish worldwide, there is a clear need for normative data and datasets that can be used to train and evaluate semi-automatic scoring algorithms to assess semantic verbal fluency performances for native Turkish speakers. However, in a systematic literature review reported in Chapter 4 (early version: Appendix A.1), I showed that there are very few normative studies for the Turkish language and that most existing normative studies are only published in Turkish, not in international literature.

1.3 Overview of Studies and Contributions

The aim of this thesis is to create a semi-automatic method for scoring Turkish SVF sequences that are assessed based on a novel, publicly available dataset of SVF sequences produced by healthy native speakers of Turkish from Türkiye and the Turkish diaspora. The main research questions and contributions are as follows.

1. **Semantic Verbal Fluency in Native Speakers of Turkish: A Systematic Review of Category Use, Scoring Metrics and Normative Data in Healthy Individuals**

   RQ: What is known about SVF performance in native speakers of Turkish with regard to semantic categories, scoring metrics, and normative data?

   Contribution: I showed that next to no normative data on Turkish SVF is available in the international literature. More advanced metrics, such as clustering and switching, have barely been reported on, and there is a lack of computational analysis tools. Despite the size of the Turkish diaspora, very little is known about the SVF performance of multilingual versus monolingual speakers of Turkish.

2. **Algorithm Validation: Semantic Verbal Fluency in Spanish-Speaking Colombian Alzheimer's Disease Patients**

RQ: How successful are the computational linguistic methods that will be adapted for Turkish in replicating a manual annotation of the internal structure of SVF sequences? To what extent can metrics derived from computational analysis distinguish between Colombian Alzheimer's disease (AD) patients suffering from a specific single gene mutation and healthy controls?

Contribution: To the best of my knowledge, this is the first study reporting a clustering/switching analysis for a SVF dataset gathered from people with familial AD resulting from a mutation in the presenilin-1 (PSEN-1) gene. I showed that a simple bigram-based method is sufficient to distinguish between people with and without AD. However, methods such as the vector space method that leverage large linguistic datasets outperform the bigram approach, which implies that bigrams are best suited for substantially under-resourced languages. The source code for both methods is available in a publicly available repository so that it can be adapted to other languages.

3. **Online Collection of Turkish Semantic Verbal Fluency Data: Benefits and Challenges**

   RQ: How can we effectively design an online data collection tool for Turkish SVF data? What are the benefits and challenges of a fully online data collection strategy?

   Contribution: In the Turkish studies examined in our systematic review study, we generally observed that researchers tended to collect first-hand data from their target group through in-person meetings. However, during the COVID-19 pandemic, lockdowns prevented in-person data collection in health care facilities. I therefore used a fully online data collection approach. Moreover, we have made our web based application publicly available and discuss lessons learned for other researchers who wish to build on our approach.

   With the dataset itself, I make three main contributions:

   (a) To the best of my knowledge, this dataset is the first Turkish SVF dataset that has been collected completely online.

   (b) To the best of my knowledge, this dataset is the first Turkish SVF dataset made available to other researchers.

   (c) The dataset includes monolingual, bilingual, and multilingual participants, which addresses a gap in studies identified in the systematic review.

4. **Analysis of Semantic Verbal Fluency in Data Collected from Turkish Speakers**

   RQ: To what extent are the computational methods introduced earlier in Chapter 5 the successful at replicating the internal structure of SVF sequences in Turkish? Can metrics derived using these methods replicate potential performance differences based on gender and multilingualism?

   Contribution: I successfully adapted my implementations of the bigram and vector space methods to Turkish and showed that they can successfully replicate manual clustering and switching analyses. I performed an in-depth comparison of the analysis results from our dataset and both international and Turkish norms in terms of different demographic elements.

1.4 Overview of Chapters

This outline of the thesis presents a brief summary of each chapter.

Chapter 2 – The Semantic Verbal Fluency Task: Background and Manual Analysis: In this chapter, the history of SVF tests and how the current standardised version was formed is outlined in chronological order. Then, we focus on methods of data collection and describe common performance measurement techniques. We examine in detail the traditional word counting methods and the Troyer manual method, acknowledging its great contribution to the field. In addition, we discuss factors that affect SVF performance, specifically focusing on demographic characteristics specifically age, gender, education, and language status. Lastly, we extend the discussion to the effects of behaviours such as physical exercise, and leisure activities on SVF.

Chapter 3 – Computational Linguistic Analysis of Semantic Verbal Fluency Data: In this chapter, we discuss extensively investigated algorithm-based methodologies for SVF which have become alternatives to traditional methods. We examine the approaches chronologically with sharing architectural details for each algorithm, considering relevant studies and their findings. Importantly, we discuss both the pros and cons of these methods, because some methods have been found to be more successful than others and have been frequently evaluated. Afterwards, we discuss the methods that we have employed in this thesis, with a step-by-step description of their implementation. While concluding this section in which the methods are explained, we also provide the most up-to-date approaches that are not yet widespread in the SVF field

but will be good candidates in the near future.

Chapter 4 – Semantic Verbal Fluency in Native Speakers of Turkish: A Systematic Review of Category Use, Scoring Metrics, and Normative Data in Healthy Individuals: In this chapter, we perform a systematic review of studies examining semantic verbal fluency in Turkish speakers. The review mainly focuses on performance evaluation metrics, widely researched categories, and norms for healthy people. Studies that included at least one healthy group were also included to identify diseases that were frequently investigated. The main purpose of this study was to provide a resource for researchers by bringing together SVF studies for individuals whose mother tongue is Turkish.

Chapter 5 – Algorithm Validation: Semantic Verbal Fluency in Spanish-Speaking Colombian Alzheimer's Disease Patients: In this chapter, we explore the use of two computational linguistic techniques to study a special case of hereditary dementia cases for Colombian Spanish speakers. We compare the results of algorithm-based methods with those derived using the Troyer method, also called the manual method, and examine how successful computational methods are in distinguishing healthy individuals and patients with dementia. Traditional word count metrics are also presented as a baseline.

Chapter 6 – Online Collection of Turkish Semantic Verbal Fluency Data: Benefits and Challenges: In this chapter, we present the Turkish SVF dataset collected entirely online through a participant-centred application. We emphasise the details of the collection steps and the difficulties encountered in this process to enable reproducibility. We also include suggestions and advice for researchers on improving data quality in online data collection. Additionally, we provide a fundamental analysis of the criteria for accepting or excluding each record, and explore how participant behaviors may be influenced by the design approach within the context of existing literature.

Chapter 7 – Analysis of Turkish Semantic Verbal Fluency Data Collected Online: In this chapter, we analyse our own Turkish SVF dataset using computational linguistics methods that are common in the field as alternatives to traditional techniques. We discuss how the results reflect the internal structure of the SVF dataset in terms of differences between demographic groups by comparing the clustering and switching components and how they differ from the results obtained through the traditional methods. Furthermore, in order to reveal the similarities and differences between languages in terms of SVF, we compare the results obtained for the Turkish language with the results of studies using similar methods for other languages.

Chapter 8 – Conclusion:In this chapter, we provide an overview of our thesis and the studies conducted. While highlighting the contributions of our work, both to the field and more specifically to research in the Turkish language, we map our findings within the framework of norms in the field of SVF. We also acknowledge the limitations of our studies. Finally, the future studies section mentions studies that could not be included in the present thesis but that could be derived from the research we have done and that would contribute to the field.

Chapter 2

The Semantic Verbal Fluency Task: Background and Manual Analysis

2.1 A Brief History of Semantic Verbal Fluency Test

The predecessor of the verbal fluency (VF) test used today is the 'word fluency factor', which is a sub-test of Thurstone's Primary Mental Abilities Battery used to determine intellectual capacity (Thurstone and Thurstone, 1938).

Bousfield and Sedgewick (1944) examined how associative thinking skills change over time based on restricted instructions and experimented extended data collection encompassing both semantic verbal fluency (e.g., bird naming, US cities, quadruped mammals) and phonemic fluency (e.g., surnames beginning with the letter S, words containing the letter M) tasks. Bousfield and Sedgewick (1944) called the procedure as a 'word association experiment' and found that the number of words produced in tasks decreases over time, with participants failing to exceed 18 minutes while completing the tasks. That study was a valuable precursor to later research. According to Ruff et al. (1996), the first revised model of VF, published by Borkowski et al. (1967) and Benton (1968), was called the 'Word Fluency test' because they were inspired by Thurstone's Test. After a while, they changed the name to 'Controlled Oral Word Association Test' (COWAT). The point is that the verbal fluency test type used by Benton and his colleagues in both studies was Phonemic verbal fluency (PVF). Isaacs and Kennie (1973) proposed an improved version of semantic verbal fluency (SVF) test using four categories: colours, animals, fruits, and towns. Since verbal fluency types have been introduced to literature at different times, ambiguities have emerged. For instance, although COWAT is another name for a PVF test, it has also been referred to as a

verbal fluency test by some researchers such as Ardila et al. (2006).

2.2 Attributes of SVF test

The standard practice in SVF data collection involves the practitioner counting as many words or names as the participant can produce between the 'start' and 'stop' instructions. Participants are timed with the aid of a stopwatch. When the participant hesitates or pauses, the practitioner can assist them with statements encouraging them to continue, but expressions that guide them or help them to remember should be avoided. Before starting the data collection process, there are two key factors of the SVF test that must be decided: the semantic categories and the duration.

2.2.1 Semantic Categories

Semantic categories are organised groups of words or names with common characteristics and are based on semantic knowledge extracted or obtained from life experiences (Kintz and Wright, 2017). Semantic cognition refers to a set of neurocognitive mechanisms involved in providing verbal and behavioural representations of semantic knowledge; these mechanisms include the processes of selection and retrieval of information (Ralph et al., 2017; Hoffman, 2018). During data collection for SVF, various semantic categories are utilized to observe changes in participant's performance. Isaacs and Kennie (1973) provides an early example of the types of categories used in measures of SVF. In this study, data were collected from older people using four categories: colours, animals, fruits, and towns. Currently, data collection is carried out in a more standardised and systematic way involving many more categories, such as supermarket items, foods, vegetables and fruits. Animal naming is the most studied category in semantic verbal fluency test (Patterson et al., 2011). Ardila et al. (2006) outlined the reasons for this preference for the animal category as follows:

- It is a category that participants can easily generate words for regardless of language, culture and demographic features.

- It is easy to evaluate and compare results for practitioners because there are many examples of studies using this category in the field.

2.2.2 Test Duration

Test duration is the amount of time participants are given to produce words in a particular task. Most studies collecting SVF data have used a test duration of 60 seconds (Hurks et al., 2010; Kim et al., 2011; Lopes et al., 2009a; Hurks, 2012; Takács et al., 2014; Zimmermann et al., 2014), meaning that results can be compared easily across equivalent studies.

However, there are studies in this field that have used a test duration of 90 seconds; this is the long version of the SVF test (Dritschel et al., 1992; Pagliarin et al., 2021; Chasles et al., 2020; Fonseca et al., 2021). Baldo et al. (2010) stated that the 90 second duration is selected to allow the researcher to observe the changes in participants' performance over a longer time period. Additionally, some studies employ the short version of the SVF test, which is applied in 30 seconds. Short version tests have mostly been implemented to show that data collected in a short period of time is sufficient to distinguish patients from healthy people. Examples of this usage are Herrera-García et al. (2019) in cognitive impairment and Kim et al. (2011) in aphasia. Since the detailed tests can present a burden to those in the advanced stages of clinical disorders, Kim et al. (2011) advocated for short versions to reduce stress on both the patient and the practitioner.

2.3 Data Collection Strategies

The increase in computer-aided approaches has led to significant changes over time, not only in data analysis but also in data collection. Here we take a look at the techniques commonly used in data collection from the past to the present.

2.3.1 Paper and Pencil Assessments

Paper and pencil assessments are traditional data collection techniques in many areas of research and are still frequently used in clinical practice (Vermeent et al., 2020). The method is applied through in-person meetings between the participant and the practitioner (Staffaroni et al., 2020). Although the practitioner typically writes down the words requested from the participants (Zhao et al., 2013; Obeso et al., 2012; Mironets et al., 2023; Pakhomov et al., 2012), there are also studies where participants write words on their paper and the practitioner is only responsible for giving instructions and timing (Scheuringer and Pletzer, 2017; Abrahams et al., 2000; Shao et al., 2014;

Butković, 2018).

The major **advantage** of this data collection technique is that it requires very few materials—just a pencil, paper, and stopwatch—and that it provides an opportunity for observation during the data collection process. Since this technique does not require any technology, it is the first candidate for rapid data collection. The technique is still frequently used in clinical scenarios, where data collection is difficult, and is especially employed with elderly patients (Sternin et al., 2019). However, the technique also has **disadvantages**. It is labour-intensive (Vermeent et al., 2020), because it requires face-to-face meetings and manual scoring and analysis, unless the SVF sequences are entered into an automatic scoring tool.

2.3.2 Audio Recordings

Audio recordings are usually obtained from face-to-face meetings during which a voice or tape recorder is used to record participants' output. It is a simple and effortless way to obtain data that can be examined later. In some studies, the recordings are transcribed verbatim, whether by having transcribers type everything they hear in the recordings word-by-word (Quaranta et al., 2019; Young et al., 2015; Pakhomov et al., 2015), or by using automatic speech recognition systems (ASR) or applications (Tröger et al., 2019; König et al., 2018a; Ayers et al., 2022; Bushnell et al., 2023).

This approach has the **advantage** of avoiding possible typos or mistakes in word order. It can also be adapted to new approaches; audio recordings allow for not only word-based analyses but also investigations of more detailed audio prosody features (e.g. pauses,intonation). This technique also has the benefits of face-to-face data collection, such as being able to give participants statements encouraging them to continue. However, when the audio needs to be manually transcribed, audio recordings also have the **disadvantage** of requiring manual scoring and analysis. When automatic speech recognition is used, the analysis can be fully computational. However, precautions should be taken when collecting data; the recording devices should be allocated specifically to the task and must not be personal devices. Since the recorded data includes participants' voices and is therefore identifiable, data confidentiality must be ensured meticulously. Audio recorders and drives containing collected data should be kept in a locker accessible only to authorities and should not be removed from the data clinic.

2.3.3 Telephone Interviews

Telephone interviews are still the main alternative to in-person meetings since the telephone is a widely used and well established communication tool. The main purpose of this data collection method is to record the conversation so that it can be examined later. In this respect, this method resembles audio recording during face-to-face meetings. Many cognitive batteries, some including verbal fluency tests, have been converted into telephone-based versions that have been validated and employed by many researchers. Validation studies have shown that data obtained by telephone interviews is promising and open to improvement (Castanho et al., 2014). In their review study, Castanho et al. (2014) summarised some of the validated telephone based version of cognitive batteries, including those that incorporate verbal fluency tests such as the Minnesota Cognitive Acuity Screen (MCAS) (Knopman et al., 2000) and the Montreal Cognitive Assessment (MoCA) (Pendlebury et al., 2013). One of the largest datasets collected via telephone, *The Canadian Longitudinal Study on Aging* (CLSA) (Raina et al., 2009), collected clinical, psychological, and demographic data from a total of 50,000 French and English speakers aged 45 to 85. Later Taler et al. (2020) used the only semantic verbal fluency data from CLSA dataset, conducting novel computer-based approaches to scoring data from around 12,000 participants. Other studies that have used telephone-based SVF tests include Marceaux et al. (2019); Bunker et al. (2017); Aiello et al. (2022); Tröger et al. (2018); Gregory et al. (2022).

2.3.4 Online Applications

With the rapid rise of computer science applications in many areas, developments in data access strategies have enabled researchers to collect data online and store data digitally in secure local or cloud storage systems. Many cognitive batteries have been converted into online versions in order to standardise test administration and accelerate the speed of data collection and analysis. The Brain Health Registry (BHR), an initiative aiming to provide large-scale data for researchers, collects data from individuals who have experienced neurocognitive disease, their relatives, and healthy participants (Nosheny et al., 2015; Weiner et al., 2018). To assess various cognitive domains such as memory, attention, and processing speed, BHR employs fully online self-administered cognitive batteries, including the Cogstate Brief Battery (Maruff et al., 2009), Lumos Labs NeuroCognitive Performance Tests (Morrison et al., 2015), and the MemTrax Memory Test (Ashford, 2005).

This method has many **advantages** compared to other collection techniques. With this method, it is possible to provide a quick automatic analysis based on samples from people with similar linguistic, cultural, and demographic characteristics (Sternin et al., 2019). Systems that can be fully self-administered (i.e. those in which participants use an application based on given instructions, without a practitioner) can easily be administered over long distances, provide convenience in terms of workload, and prevent time losses in studies (Staffaroni et al., 2020).

On the other hand, there are some **disadvantages** of online data collection methods. From the perspective of the data collector, the need to design and establish web-based systems necessitates the involvement of experts from different fields. Also, web hosting (domain) services must be provided and maintained, which can be costly. In a comparison of different data collection techniques, Namey et al. (2020) revealed that online video-based data collection methods are the most costly, mainly due to platform fees. From the participant's perspective, online methods have certain technological requirements, such as a computer or phone, internet connection, and additional tools such as video and audio recording devices, depending on the nature of the collected data. Furthermore, participants need to be able to use technology equipment easily, and elderly individuals may not comfortable in this regard. Studies have shown that internet usage among the elderly is 42.7% in the United States (Gell et al., 2015) and 49% in European countries (averaged across 17 countries) (König et al., 2018b). In their study based on the 2017 data from the Turkish Statistical Institute (TUIK), Aytuna and Çapraz (2018) indicated that the internet usage rate among individuals aged 65 and older in Türkiye was only 5%. Although these numbers may increase over time, this situation may mean that online collected data is affected by *sampling bias*. This topic will be discussed in-depth in Chapter 6.

To the best of our knowledge, web-based online data collection has not been widely used in the field of verbal fluency. We only came across one study, conducted by Cho et al. (2021), which collected F-letter fluency data from university students who received course credit as a compensation. As far as we know, the BHR does not include a verbal fluency component.

2.4 Measuring Performance

The output of an SVF test is the list of words produced in a given category, such as animal names or supermarket items. This section will review two most commonly

accepted approaches to assessing participant performance: (1) word-count metrics and (2) clustering-switching. Moreover, we will provide an overview of additional attributes, besides lexical and semantic features, that can help to sensitise in-depth SVF analysis including factors such as time variance and prosody.

2.4.1 Traditional Methods: Word-Count Metrics

In an SVF test, each word produced by the participant in the predetermined amount of time is considered to determine their performance. The participant is expected to produce words in the given category in accordance with the instructions, and the words produced can be scored as follows.

Total word count: The most commonly used scoring metric for SVF is the total number of correct words belonging to the given category produced within the time limit (Ardila et al., 2006). The term 'correct' refers to the words produced that belong to the desired category. For example, 'cat', 'dog', 'lion', and 'eagle' would all be correct words given the animal category. Some studies consider animal classes or supracategory names (e.g. birds, fishes, insects) as valid names of animals (Stokholm et al., 2013; Rodríguez-Lorenzana et al., 2020), while some exclude them and only count specific animal names (e.g. sparrow, salmon, butterfly) (Gocer March and Pattison, 2006; Raoux et al., 2008).

Sometimes, participants repeat words or use both singular and plural forms of a word. If a participant produces both the singular and plural, like 'dog' and 'dogs', the plural word is converted into its singular version, and the output includes one repetition of 'dog'. Many studies eliminate repetitions and count the remaining words, concentrating on the **total number of correct unique words** (Rodríguez-Lorenzana et al., 2020; Raoux et al., 2008; Gocer March and Pattison, 2006; Kavé, 2005; Kosmidis et al., 2004b). However, Tröger et al. (2019) emphasised the importance of repetitions and removed them only if they occurred consecutively, leaving other repetitions unaltered.

There are also studies that have reported the number of words produced in the first, second, third, and fourth quarters of the 60 second timeframe to show how word production changes over time. We will discuss time intervals in Section 2.4.3.2 as other features.

Errors: Errors provide a qualitative measure of SVF performance and are divided into two sub-types: category violations and perseverations (Raboutet et al., 2010).

1. **Category violations:** Also referred to as intrusions, category violations are additional words that do not belong to the desired category (Raskin and Rearick, 1996a). For the animal category, this would include any word that is not the name of an animal.

2. **Perseverations:** Perseveration refers to the repetition of words and is divided into three types (Sandson and Albert, 1984; Goldberg, 1986; Sandson and Albert, 1987; Helm-Estabrooks et al., 1998; Pekkala et al., 2008):

   (a) *Recurrent:* Repeated words separated by other words. Example sequence: cat, dog, fox, cat.

   (b) *Continuous:* Repeated production of a single word. Example sequence: cat, cat, cat.

   (c) *Stuck-in-set:* Repetition of words from a previous category. This would be exemplified by a situation in which the researcher asks first for the animal category and then later for supermarket items, but the participant continues to list animal names. Gillen and Rubio (2016) explained that this occurs due to the previous task not being completed and the new task not being activated.

   The commonality is that all repeated words tend to be classified as perseverations (Galaverna et al., 2016; Henry and Phillips, 2006), regardless of their location or semantic integrity. Table 2.1 shows a hypothetical SVF sequence for the animal category. In this hypothetical scenario, a total of 12 words were produced by the participant. These comprise the raw data before pre-processing. According to traditional metrics, there are 9 unique words, 1 category violation, and 2 perseverations.

2.4.2 Clustering and Switching

While the commonly used evaluation criteria rely on traditional word count techniques, they are unable to provide insights into the relationships between words or the internal structure of SVF sequences. Two features, clustering and switching, have been suggested by researchers to overcome this limitation and have been investigated frequently. The clustering theory put forward by Bousfield (1953) suggests that there are

Condition	Output
Initial (raw) SVF performance in Animal category	cat, dog, lion, camel, dogs, cats, fox, water, horses, goats, sheep, cow
After converting plural to singular	cat, dog, lion, camel, dog, cat, fox, water, horse, goat, sheep, cow
Correct unique words	cat, dog, lion, camel, fox, horse, goat, sheep, cow
Category violations	water
Perseverations	dog, cat

Table 2.1: Traditional metrics: unique words (in light pink background), category violation (in blue background), and perseverations (in green background) exemplified through a pseudo SVF output belongs to hypothetical participant.

semantic relationships between consecutive words produced within a particular category. In other words, if people are counting animal names, they tend to say bird species one after the other or fishes consecutively. The theory is based on the assumption that clustering is related to the organisation of thought and recalling from the memory. Bousfield (1953) conducted their research with four different categories which are animals, human names, professions, and vegetables. The aim of the study was to group consecutive words that were related somehow, so they manually created clusters (subgroups) from the sequences the participants produced (e.g. birds, fishes for animal category). Many researchers have applied this technique by creating and expanding subgroups based on the category used, on different study populations, including healthy participants (Gruenewald and Lockhead, 1980) and patients with Parkinson's disease (Raskin et al., 1992a), schizophrenia (Allen et al., 1993), and Alzheimer's disease (Binetti et al., 1995).

Although the clustering technique was widely accepted in the field, there was no common protocol to follow. To that end, Troyer et al. (1997) created an animal taxonomy by dividing animal species into three main groups and twenty-two subgroups that are semantically related. The main and subgroups of the Troyer Taxonomy are given below. The full Troyer taxonomy list, including animal names, is available in the supplementary documents (see Appendix A.6).

- **Living Environment:** Africa, Australia, Arctic/Far North, Farm, North America, Water
- **Human Use:** Beasts of burden, Fur, Pets
- **Zoological Categories:** Bird, Bovine, Canine, Deer, Feline, Fish, Insect, Insectivores, Primate, Rabbit, Reptile/Amphibian, Rodent, Weasel

Building on the taxonomy outlined above, Troyer et al. (1997) proposed a more systematic approach to creating animal groups, called the clustering component. They also introduced a new component called switching. In Figure 2.1, a hypothetical SVF sequence is divided into four clusters and three switches according to the Troyer taxonomy protocols. The definitions of Troyer's clustering and switching are as follows:

- **Cluster:** A cluster is made up of consecutive words produced by the participant that belong to the same subgroup. In Figure 2.1, the first three animal names belong to the 'North America' group, which is called Cluster-1.
- **Switch:** If words adjacent to the produced names belong to the different subgroups, a break occurs between clusters. The tendency to create a new cluster is called switching. As an example, in Figure 2.1, the naming 'Beasts of burden' animals after 'North America' animals caused a switch.

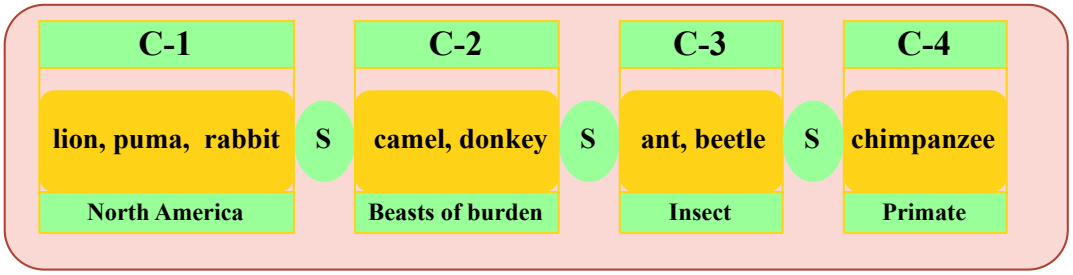


Figure 2.1: Clusters and switches created according to Troyer's animal taxonomy

Some animals belong to more than one group in the taxonomy. For example, 'dog' and 'fox' belong to both 'canine' and 'fur' groups. Therefore, certain protocol rules for the creation of cluster and switch components were determined by Troyer et al. (1997). The main rules are as follows:

- If two consecutive words belong to the same group, they create a cluster. Additional related words add to this cluster, until a word is produced that breaks the rule.

- If the cluster is unclear because two adjacent words both belong to more than one group, the third word is checked. The goal is to create the largest cluster with consecutive words.
- If consecutive words do not belong to a group, a cluster consists of one word.
- If the given word is not suitable for the nature of the test, this is a category violation [1]. If a participant says a word more than once, these are called repetitions. Category violations and repetitions are not excluded.

Since the Troyer method has a clear protocol, it is easily adaptable and widely used. As a result, many studies have accepted Troyer's animal taxonomy as a gold standard for comparisons between different methods of animal naming. Studies have used this standard to study dementia (Paula et al., 2018), mild cognitive impairment (MCI) (Oh et al., 2019), Alzheimer's and Parkinson's diseases (Tröster et al., 1998), and the brain health of professional fighters (Ryan et al., 2013), among others.

The Troyer method consists of predefined lists containing a limited number of animal names, which causes some drawbacks. Determining appropriate groups for produced names that are not included in the existing taxonomy is challenging, and new groups are needed. Furthermore, the fact that some animals belong to more than one group causes ambiguity, so it is hard to create clusters properly, as emphasised by Woods et al. (2016a). These difficulties additionally necessitate manual annotation, including translation of the method into target languages, making use of the for the Troyer taxonomy is labour intensive.

2.4.3 Other Methods

2.4.3.1 Utilisation of Non-Lexical Features

Total word count and errors rooted in lexical analysis have long been the bases of SVF assessment. Although this thesis focuses on word-based metrics (clustering and switching), which are the evaluation features commonly used to measure SVF performance, we should note that beyond lexical attributes, **prosody** features (e.g. pauses, stress, intonation) are also often explored to delve into the acoustic landscape of speech. Common in typical conversation, nonverbal prosodic details are elements that provide the listener with information about the person's mood, feelings, and attitude (Mitchell

[1] Troyer et al. (1997) used the term 'error' rather than category violation. We have changed it here as errors is an umbrella term which includes both category violations and perseverations.

and Ross, 2013). However, different vocal characteristics are associated with neurocognitive conditions. For example, abnormal speech rate is observed in bipolar patients (Raucher-Chéné et al., 2017), mono-pitch and reduced stress in Parkinson patients (Jones, 2009), and flattened speech lacking emotion in Alzheimer's patients (Tosto et al., 2011).

Troyer et al. (1997) asserted that clustering is linked to semantic retrieval and involves temporal lobe activity, whereas switching relies on executive functions associated with frontal lobe activity. However, Mayr (2002) pointed out that the switching component in the Troyer method is ambiguous. Switching refers to transitioning between semantic clusters, but if an individual struggles to produce words within a semantic cluster, a decrease in the number of switches may be observed. In other words, switching will change not only when executive functions are involved but also when a person encounters difficulty in retrieving semantic information. Furthermore, the more time a person spends within a cluster, the more their ability to create new clusters and switch will be affected. Mayr (2002) emphasised that the evaluation of the clustering and switching components independently of the time factor is a shortcoming of the method and proposed two time courses for assessment: within-cluster retrieval duration and switch duration. Therefore, it is recommended to measure reaction times in addition to evaluating clustering and switching skills.

Mayr's time course–focused method, proposed in Mayr and Kliegl (2000), was tested by McDowd et al. (2011). They highlighted that individuals with Alzheimer's disease exhibit the slowest initial responses compared to participants with Parkinson's disease and healthy participants. This finding supports Mayr's theory, suggesting that time features can provide further insights into inter-group differences. Many other studies have suggested that by incorporating information processing and retrieval times, as proposed by Mayr, it may be possible to obtain more comprehensive results regarding clustering and switching components (Fossati et al., 2003; Unsworth et al., 2011; Da Silva et al., 2004; Gordon et al., 2018).

Wolters et al. (2016) examined a dataset obtained from young healthy native Korean speakers in terms of prosody characteristics and found that the number of pauses and the articulation rate were correlated with the number of switches. Another study conducted on French speakers by Weiner et al. (2021) used SVF-derived prosodic (e.g. pause) and vocal features (e.g. voice frequency) to train their classification algorithms and they successfully differentiate mixed episodes of bipolar disorders with high scores. Amunts et al. (2021) trained a machine learning algorithm on both prosodic fea-

tures and traditional word-based features from SVF data obtained from healthy German speakers and found that it could successfully predict the executive function scores of unseen people. Overall, these and other studies have shown that prosodic features can either complement word count features, enhancing the depth of fine-grained analysis, or stand alone with an informativeness comparable to that of lexical features.

2.4.3.2 Time Intervals

We previously discussed the use of word count (including errors) and word grouping strategies for performance evaluation, as well as the combination of non-word-based prosodic features. In addition to these features, some studies have broken down total word production time into blocks to analyse individuals' performance during particular time intervals, typically using 15-second blocks. Since time intervals are not the primary focus of this thesis, we will provide a concise overview of studies that use time blocks and their findings.

Pagliarin et al. (2021) conducted a comparative analysis of SVF performances from healthy individuals and those with various forms of brain damage, with and without aphasia. They examined data collected over a period of 90 seconds, which they then divided into 30-second intervals. Healthy individuals displayed the highest performance and individuals with left hemisphere damage with aphasia the lowest performance. All groups performed much better in the first 30 seconds compared to the other time blocks. Pagliarin et al. (2021) explained this pattern by suggesting that frequently used words are easily accessible at the beginning, but as time goes on, there is an increasing cognitive load on the regions responsible for executive function. Raboutet et al. (2010) confirmed this hypothesis in a comparison of healthy individuals' performance in the first and second 30-second segments. Two additional studies, Kim et al. (2011) on aphasic and non-aphasic stroke patients, and Herrera-García et al. (2019) on patients with cognitive impairment and dementia, have demonstrated that longer test duration in patient groups can increase participant frustration levels. Furthermore, they found that using either a 30-second or 60-second test duration did not significantly impact the ability to differentiate between patient groups; therefore, shorter assessment periods were recommended for groups facing challenges with data collection. In conclusion, the ability to recall words is initially relatively high in both healthy individuals and patient groups but decreases throughout the task. Collecting data for a shorter period can accelerate the process and reduce pressure on individuals with advanced neurocognitive disorders. However, this could compromise in-depth analysis and po-

tentially lead to information gaps in detecting early-stage cognitive declines, since a longer test period places greater demands on executive functions and therefore gives better insights.

2.5 Demographic Determinants on Performance

Although SVF outcomes can be analysed with any selected scoring technique, the results need to be compared with population norms derived from normative data. **Normative data** are data obtained from a particular group of people at a particular period of time that illustrate what is usual for the target population (O'connor, 1990). The data are used as a reference point to provide a benchmark for other studies that investigate the same population, so they include randomly selected representative individuals covering different demographics (Campbell, 2021). The most commonly investigated demographic characteristics are gender, age, and education, but socioeconomic status is also sometimes investigated through the collection of information about employment status or income level. Language status—whether an individual is monolingual, bilingual or multilingual—is another determinant widely investigated in cognitive studies. This section will focus on commonly used demographics and their effects on SVF performance.

2.5.1 Education

Education has a positive effect on cognitive function, and it is expected that a higher level of education would be associated with increased performance (Lövdén et al., 2020). Moreover, according to Guerra-Carrillo et al. (2017), higher education has long-lasting effects and can help better cognitive abilities in older ages. Several studies have examined the effect of education on SVF test performance in order to measure cognitive abilities. Education can be quantified using levels (primary, secondary, university, etc.) (Mathuranath et al., 2003) or in intervals of years (0-2, 3-5, 6-9, etc.) (Benito-Cuadrado et al., 2002b). Due to differences in education systems among countries, the appropriate education levels and year intervals vary between studies. Moreover, compulsory education differs between countries and similar levels may start earlier or end later in different countries, making it hard to compare them. For instance, the age range for compulsory education is 5-16 years in the United States of America (USA) (Wood, 2006) and the United Kingdom (UK) (Millar, 2002), 6-14 years in In-

dia (Jha and Parvati, 2014), and 6-17 years in Türkiye (Dayıoğlu and Kırdar, 2022). For these reasons, there is usually no standard education level, and researchers collect data according to the education systems of the countries in which the study is conducted. In order to overcome this situation, (Ostrosky-Solis et al., 2007) compared SVF outcomes between different countries speaking the Spanish language, using years of education rather than levels.

In Terms of Word Count Evaluation in SVF: Studies conducted in different languages and with different categories have shown that education has a significant positive effect on SVF performance. In the light of the normative data provided by later studies, they confirmed great effect of education and added that the production of words increases as the level of education increases. This was demonstrated for the animal category in Portuguese (Brazil) by Esteves et al. (2015), English (Canada) by Tombaugh et al. (1999), and Turkish (Istanbul, Türkiye) by İlkmen and Büyükişcan (2022); and in multiple categories (names, supermarket objects, animals, kitchen objects, food, transports, and clothes) in Portuguese (Portugal) by Nogueira et al. (2016).

Apart from the number of years of education received, **literacy** is also one of the factors examined in relation to SVF performance. Da Silva et al. (2004) conducted an experiment comparing illiterate and literate healthy female Portuguese speakers above 60 years old; there was no age difference between groups. They observed a significant difference in favour of literate individuals for the animal category but not for supermarket items. In a study conducted on healthy elderly individuals in Brazil, Fichman et al. (2009) concluded that education has a significant positive effect on naming in the animal category and that even people that have 1-4 years of education had much better performance compared to illiterate participants. These results suggest that while each additional level of education has a positive effect on SVF, the greatest difference is observed between literate and illiterate individuals. Nielsen and Waldemar (2016) examined the effect of illiteracy on SVF performance in the animal and supermarket categories among Turkish speakers; confirming the results Da Silva et al. (2004), they found no difference between groups in supermarket products and significantly better performance among the literate group in the animal category. Nielsen and Waldemar (2016) explained the difference between these two categories, arguing that supermarket products create a vocabulary that people can experience in daily life through shopping, resulting in similar knowledge regardless of literacy. Significant differences are observed in the animal category, however, because literate people extend their animal knowledge by reading printed materials (e.g. books, newspapers), while illiterate peo-

ple have a more limited number of animals that they know from their life experiences, such as farm animals and pets.

In addition to studies on healthy individuals, the effect of education has also been investigated in people with cognitive disorders. A study of Japanese-speaking individuals conducted by Kawano et al. (2010) evaluated patients with MCI and probable AD on animal naming, and found that in both groups, those with nine or more years of education scored higher than those with less than nine years. Another study was conducted on Brazilian Portuguese speakers by Radanovic et al. (2009), the study compared two different education levels in the animal category in healthy individuals, patients with MCI and probable AD. In healthy and MCI individuals, those with 8 years or more education performed significantly better than those with 0-4 years of education, while the difference between the education groups in AD patients was low to be neglected (in favor of low education). However, Mirandez et al. (2017) carried out a study of the effect of education on MCI patients who speak Brazilian Portuguese and found no difference between different education levels. In a study on Chinese speakers comparing healthy individuals and patients with AD, Mok et al. (2004) observed a positive effect of education in healthy individuals but similar scores at different education levels among AD patients. Furthermore, the performance of literate AD patients was higher than those who were illiterate. In other words, although illiteracy and literacy created a performance difference among AD patients, no effect of formal education on performance was observed. In brief, literacy has a significant effect on both healthy people and patients with cognitive deficits, but the effect of further education on progressive dementia is unclear. Also, the effects of education on mild cognitive impairment is still controversial and needs substantial future research.

A limited number of studies have also investigated how education affects error types, but this has not been widely studied. Da Silva et al. (2004) found no difference in perseverations or category violations between illiterate and literate Portuguese speakers in supermarket items and animal naming. López-Higes et al. (2022) studied Spanish speakers, people from those who have no formal education to higher education level, and concluded that education has no effect on any error types in four categories: animals, fruits, kitchen tools, and clothes. However, Kosmidis et al. (2004b) studied individuals with1-21 years of education in three levels and pointed out that more education was associated with increased perseverations in Greek.

In Terms of Clustering and Switching Evaluation in SVF: Beyond the number of words produced, education gives an impression about the strategy of group-

ing these words. The studies conducted on healthy participants on various languages report that as the education level increases, the number of switches and clusters increases, in this case the mean cluster sizes decrease; Pereira et al. (2018) in Brazilian Portuguese with clothes, Brucki and Rocha (2004) in Brazilian Portuguese with animals, Troyer (2000) in English with animals and supermarket items, Kosmidis et al. (2004b) in Greek with animals, fruits, and objects. In other words, educated people tend to change subgroups (for animals: farm, pets, water, birds, etc.) more frequently. Although Brucki and Rocha (2004) state that the number of switches and clusters increases gradually with education years that participants spend, they also emphasise that no difference in significance level between levels that are close to each other: for example, between 9-11 and 11+ years or between illiterate and 1-4 years of education. On the other hand, Da Silva et al. (2004) pointed out that there is a significant difference in favor of literate in both clustering and switching numbers between literate and illiterate groups.

2.5.2 Age

Aging is characterised by cellular and physiological degradation in the organism, and as it progresses, the individual becomes weaker and more susceptible to diseases (Rose, 1991; Booth and Brunet, 2016). As well as physical deterioration, aging causes cognitive, functional and social impairments (Bettio et al., 2017). Cognitive decline is associated with slow processing speed, difficulty in reasoning with existing knowledge, poor memory, and impaired executive functions (e.g. planning, organising, attention) (Deary et al., 2009). These impairments are closely related to the changes in the structure of the brain by age: (a) decreasing cortical gray matter density and white matter volume cause dysfunctions on biochemical transmission between neurons (Broglio et al., 2012), (b) atrophy in the hippocampus leads to difficulty maintaining the episodic memories such as experiences, events, emotions etc. (Bettio et al., 2017). Although these limited physical changes in the brain are part of normal aging and create performance differences between young and old people, advanced structural loss is associated with neurocognitive disorders cause severe loss of ability to maintain daily tasks compared to normal elderly. In this case, the question arises, what are the main differences between normal (typical) and pathological aging? Normal aging, or age-associated cognitive decline in other name, effects memory but pathological aging (i.e. dementia, Alzheimer's, mild cognitive impairment) also intervenes the social or

occupational function (Levy et al., 1994; Deary et al., 2009). Additionally, normal aging reduces little in verbal ability, numerical ability, and world knowledge but more in memory, executive functions, processing speed and reasoning (Fjell et al., 2014; Deary et al., 2009). Therefore, if deterioration effects severely in both former and latter abilities, neurocognitive diseases can be investigated.

Aging is one of the most influential factors on cognitive abilities like education, some studies suggest that aging has a greater effect (Stokholm et al., 2013), while others assert education (Zimmermann et al., 2014; Moraes et al., 2013). The elderly not only exhibit lower performance than the young in executive functions but also, pathological conditions expand the gap between healthy individuals and the patients who diagnosed with cognitive disorders. In this respect, deeper insights will be provided by examining the effects of aging on SVF performance from two perspectives: how older people differ from younger individuals, and (2) neurocognitive disorders. We will discuss these two point in the following in terms of two aspects: word count and clustering and switching features.

In Terms of Word Count Evaluation in SVF: Memory is associated with previous life experiences and older individuals possess greater knowledge than younger people. However, elderly individuals exhibit poor SVF performance because aging affects cognitive abilities and makes it difficult to access semantic information (Perlmutter, 1978). Many studies examining various semantic categories and languages in healthy individuals confirm that older people produce fewer words than younger people in SVF tasks; this was seen in Malayalam (Mathuranath et al., 2003) and in Spanish (Benito-Cuadrado et al., 2002b; Iñesta et al., 2022) for animals, in European Portuguese (Nogueira et al., 2016) for names, supermarket, kitchen objects, food, clothes, in Dutch Van Der Elst et al. (2006a) for profession and animals, in Lebanese (Jebahi et al., 2022) for accessories, animals, electronics, fruits, kitchen utensils, natural objects, professions, and sports. Benito-Cuadrado et al. (2002b) were stratified participants in six age groups between 18 and 76+ and highlighted that for every seven years of age, the SVF score was reduced by one point.

Studies have indicated that the onset of cognitive decline varies in healthy individuals. However, the most significant declines are recorded above the age of 70 (Aartsen et al., 2002; Baltes et al., 1995). As a result, the differences among groups aged above 70 tend to be less distinguishable. Haugrud et al. (2010) confirmed this argument in an experiment eliciting words in the animal category. They categorised their English speaking participants into four different age groups and found a significant difference

between the old-elderly group ($\overline{X}_{age} = 79.9$) and the two younger groups, middle-aged ($\overline{X}_{age} = 55$) and young ($\overline{X}_{age} = 28$), but no difference between the old-elderly group and the young-elderly group($\overline{X}_{age} = 70$). The findings of Tomer and Levin (1993), which used both animal and food categories, support this. Furthermore, Tombaugh et al. (1999), which also used the animal category and dividing participants into three age groups (16–59, 60–79, and 80–95 years), found that the most significant decline was found in the 60–79 age group.

While reduced performance of SVF is typically observed in healthy elderly individuals compared to young people, pathological conditions make performance worse. Studies have found a significantly lower total word count among neurocognitive patients than age- and education-matched healthy comparisons. This has been demonstrated for the animal category in English patients above the age of 70 with Parkinson's disease and dementia by Piatt et al. (1999) and French participants around the age of 65 with frontotemporal dementia and semantic dementia by Laisney et al. (2009). Similar results have been obtained in English patients above 70 years old with Alzheimer-type dementia using tasks eliciting animals, fruits, vegetables, first names, and supermarket items by Monsch et al. (1992) and just animals by Haugrud et al. (2010). In summary, while aging leads to a decline in performance, the literature highlights notable distinctions between healthy individuals and those with cognitive impairments.

In Terms of Clustering and Switching Evaluation in SVF: Although the total number of words gives information about the performance of participants, the clustering and switching method brings a holistic perspective that can extract the internal structure of SVF sequences. Troyer and colleagues established a benchmark in the field through their proposed animal taxonomy, categorising the generated words into groups. Troyer et al. (1997) divided healthy individuals into two groups: young ($\overline{X}_{age} = 22.3$) and old ($\overline{X}_{age} = 73.3$). They found no group differences in cluster size but older participants create less switches than younger group. If two groups create similar sized clusters and the elderly people ability to shift between clusters (switch) less frequently, this result is associated with the production of lower number of words in the elderly. Troyer (2000) confirms their previous results with English normative data. Later studies have supported Troyer's finding through comparisons of healthy young and old people in terms of clustering and switching in different languages: in English by Lanting et al. (2009), in Greek by Kosmidis et al. (2004b), in Brazilian Portuguese by Pereira et al. (2018); Brucki and Rocha (2004).

2.5.3 Gender

Gender is one demographic factor that has been widely researched when examining cognitive differences. Research on the binary gender identities (women and men) is widespread, and there are also studies on individuals within the LGBT (lesbian, gay, bisexual, transgender) spectrum in the field of SVF (Rahman et al., 2003; Rahman and Wilson, 2003; Rahman et al., 2005; Kheloui et al., 2021). It is important to highlight that some studies have discussed in detail the potential risk factors faced by the LGBT community (for instance, experiencing discrimination causes stress, anxiety, and depression) that increase the likelihood of cognitive decline and their effects on individuals' social, emotional, physical, and mental domains (Van Wagenen et al., 2013; Flatt et al., 2018; Correro and Nielson, 2020).

Research has indicated that cognitive abilities differ depending on gonadal hormones (estrogen, progesterone and testosterone); therefore, women have an advantage in verbal abilities, while males are stronger in visual-spatial abilities (e.g. mathematical skills) (Huang et al., 2015; Upadhayay and Guragain, 2014; Vidal et al., 2006; Vlachos et al., 2003; Herlitz et al., 1997). Verbal abilities encompass a range of skills, such as vocabulary, reading comprehension, analogies, and speech production, that can be assessed through various sub-tests. In their meta-analysis, Hyde and Linn (1988) analysed 165 studies that had researched gender differences in verbal ability and found that, while the mean score across all sub-skills was slightly higher for women than men, the effect size was small. As a result, Hyde and Linn (1988) argued that there is no observable difference between genders in terms of verbal abilities. While the SVF test primarily serves as a clinical assessment tool for evaluating executive functions, it is also considered one of the tests used to assess verbal abilities, particularly in terms of vocabulary knowledge and word retrieval (Shao et al., 2014). To this end, we will examine the effects of gender differences on SVF in terms of total word count and clustering and switching strategies in the light of previous studies.

In Terms of Word Count Evaluation in SVF: In an SVF test, most of the normative studies investigating different languages and categories have revealed no performance difference between men and women in terms of total word count. This was the case for animals in German (Weiss et al., 2006), Arabic (Saudi Arabian sample) (Khalil, 2010b), and Lebanese (Jebahi et al., 2022); for animals, fruits, kitchen tools, and clothes in Spanish (López-Higes et al., 2022); for animals and fruits in Italian (Zarino et al., 2014); for animals, methods of transport, and verbs in European

Portuguese (Nogueira et al., 2016); and for fruits in Polish (Sokołowski et al., 2020). However, there are also categories in which performance differences have been observed between genders: men achieved better scores in brands of cars (Zarino et al., 2014) and sports and tools (Jebahi et al., 2022), but women outperformed men in accessories and vegetables (Jebahi et al., 2022) and names, supermarket, kitchen tools, food, and clothes (Nogueira et al., 2016). Based on previous studies, no gender differences exist in categories like animals and fruits, but there are conflicting results for the kitchen tools, food, and clothes categories. Nogueira et al. (2016) found that women performed better in these categories, whereas López-Higes et al. (2022) did not observe a gender difference. Capitani et al. (1999) emphasised that gender differences may be related to retrieval strategies based on individuals' experiences, habits and occupations, rather than being an indicator of better semantic memory. As a result, the general consensus is that there is no difference in SVF performance between males and females in widely used categories (e.g. animals) Gawda and Szepietowska (2013b); Tombaugh et al. (1999); Mathuranath et al. (2003); Da Silva et al. (2004).

In Terms of Clustering and Switching Evaluation in SVF: Numerous studies focused on word grouping strategies in different languages among healthy individuals, have suggested that there are no significant differences between genders (female and male) in terms of the number of switches or the average cluster size. This has been shown for animals in Brazilian Portuguese by Brucki and Rocha (2004), English by Troyer et al. (1997), German by Weiss et al. (2006), and Turkish by İlkmen and Büyükişcan (2022); for animals, fruits, and sharp objects in Polish by Sokołowski et al. (2020); and for ten different categories in Austrian German by Scheuringer et al. (2017). However, Lanting et al. (2009) reported that within the animal category in English, females had a higher number of switches than males and males tended to form larger clusters than females, revealing a significant difference between genders in both features. In a study conducted in Greek, Kosmidis et al. (2004b) observed no difference between genders in the categories of animals and objects, but found that women tended to make larger clusters than men in the fruit category. Overall, most studies have showed no gender differences in terms of clustering and switching strategies but there are some studies that found different grouping strategies between men and women.

2.5.4 Language Status

Language is a vital component of culture and a symbolic expression reflecting the history, thought patterns, and lifestyle of the community it is associated with (Jiang, 2000). It is a common belief that cultural differences could impact the outcomes of working memory assessments, especially through the use of different semantic strategies to store long term knowledge (Ismatullina et al., 2014). For this reason, cognitive tests are adapted to different languages and cultures, and the results are compared to normative studies, which are carried out on healthy individuals to determine the patterns and standards of target population.

Monolingualism: Numerous cross-language studies have investigated the impact of language differences on SVF performance. Pekkala et al. (2009) compared Finnish and American-English monolinguals in two categories, animals and clothes, and revealed no differences in the total number of words between groups. However, the study emphasised that while Finnish-speaking people included mostly farm animals in their animal lists, Americans tended to focus on zoo animals. This suggests that the semantic relationship organisation may differ between Finnish and American-English speakers.Similarly, Rosselli et al. (2002) found no difference in the number of animal words produced by American-English and Spanish monolinguals, but they demonstrated that Spanish speakers tended to provide more examples of birds and insects and English speakers more often mentioned wild animals. Additionally, Kempler et al. (1998) investigated various ethnic immigrant groups in the US, including Chinese, Spanish, and Vietnamese speakers, as well as native English speakers. They focused on individuals who used their native language in their daily lives and while watching TV or listening to the radio. The Hispanic group produced significantly fewer animal names compared to speakers of other languages, while Vietnamese participants achieved the highest word count. The observation highlighted by the authors is that in the Vietnamese language, animal names typically consist of a single syllable, while in Spanish, animal names with two or three syllables are common. This difference might have directly influenced the total number of words that could be produced within a short period. To this end, although results obtained from other languages are generally useful for understanding the literature, normative studies and other studies in the target language should to be used as a benchmark. For this reason, in this thesis, the language in which the results are obtained is often stated when giving examples.

Bilingualism / Multilingualism: Another topic that researchers frequently focus

on is the effects of actively speaking two (bilingualism) or more languages (multilingualism) on recalling information from long-term memory. Various language ability surveys have been developed to measure participants' proficiency in each of their languages, such as the Language Experience and Proficiency Questionnaire (LEAP-Q) (Marian et al., 2007) and the Language and Social Background Questionnaire (LSBQ) (Anderson et al., 2018; Luk and Bialystok, 2013).

The effects of bilingualism have been extensively studied by researchers in the field of SVF and results show that bilinguals perform poorly compared to monolinguals in terms of total word count. Bialystok et al. (2008a) conducted a study comparing English monolinguals with individuals who speak another language besides English in their daily lives, and found that bilinguals generated significantly fewer words than monolinguals in the animal category. The outcome was verified by Gollan et al. (2002) in a study involving Spanish–English bilinguals and English monolinguals across twelve different semantic categories. Rosselli et al. (2000a) conducted an experiment in the United States and compared Spanish-English bilinguals to monolinguals with English and Spanish. The bilingual group was tested in two languages, and the results indicated that bilinguals produced significantly fewer words in both English and Spanish when tested on the categories of animals and fruits. In a subsequent study with the same sample, Rosselli et al. (2002) explored semantic associations and found that bilinguals formed significantly more clusters in Spanish than in English. Sandoval et al. (2010) compared English monolinguals with English-Spanish bilinguals (English-dominant) in 15 different categories, finding that bilinguals presented lower category fluency performance in total word produced (mean score of all categories) than monolinguals, as previous studies had reported.

However, there are studies that show no significant difference between bilingual and monolingual groups. Luo et al. (2010) compared English monolinguals to bilinguals who speak languages other than English (18 different languages were included) in terms of naming performance in the categories of clothing and girls' names; their results showed no significant differences between groups in both categories. In a study conducted by Taler et al. (2013), no performance differences were observed in animal naming between English monolinguals and English–French bilingual participants.

In terms of clustering and switching components, research findings from a limited number of studies can be found as follows. In their study, Patra et al. (2020) observed that Bengali–English bilinguals made more switches than English monolinguals with no significant differences for total word counts, therefore this can be a strategy em-

ployed by bilinguals to enhance word production. In another study conducted on Farsi monolinguals and Farsi-Balochi bilinguals, Mardani et al. (2020) also found a similar word count between groups, but significantly more switches in the bilingual group.

A comprehensive systematic review of 33 studies investigating the effect of bilingualism on SVF task performance found no evidence of a bilingual advantage on any SVF indices (Giovannoli et al., 2023). According to Sandoval et al. (2010), the primary reason for the bilingual disadvantage is unclear, although several alternative explanations have been proposed. The most likely reason is a delay in response times when retrieving relevant words from semantic memory (Sandoval et al., 2010). This delay occurs because participants may recall the same word in both the target and non-target languages, leading to a delay in suppressing the non-target language (Sandoval et al., 2010; Mueller Gathercole et al., 2010; Marsh et al., 2019). Rosselli et al. (2000a) highlighted that semantic categories require the retrieval of frequent words that are more basic and generally known in both languages, resulting in increased linguistic intervention and negatively impacting the total number of words produced. Beyond cross-language interference, an alternative hypothesis suggests that, despite bilinguals having a broader vocabulary compared to monolinguals, they may not know the translations of certain words in both languages (Sandoval et al., 2010), and cultural differences may result in the absence of equivalents for some words, such as local animals or fruits (Portocarrero et al., 2007).

2.6 Other Factors

Demographic variables have been extensively researched as factors that fundamentally impact cognitive abilities in general, and SVF performance in particular. However, the existence of individuals who exhibit superior performance relative to others with similar demographic characteristics, such as age and education level, has led researchers to explore alternative factors. Research has indicated that certain habits and interests can have beneficial effects on cognitive abilities. This section will briefly discuss additional factors that impact SVF performance.

Physical exercise: It is known that exercise contributes to the improvement of neurotrophic factors in the brain; these factors are proteins that support the growth and survival of neurons and have positive effects on executive functions and memory (Zimmer et al., 2018). Exercise also increases the flow of oxygen to the brain by accelerating blood flow and supports the areas of the brain relevant to executing cognitive

tasks (Alkadhi, 2018). Studies examining different exercise techniques have indicated positive effects of exercise on SVF performance. Nocera et al. (2017, 2020) conducted a 12-week pre- versus post-intervention study looking at elderly individuals participating in spin cycling and non-aerobic exercises (stretching and balance) and found that the spin group showed improved SVF performance (for the animal category) compared to the non-aerobic group. In another 12-week follow-up study on elderly individuals, Welford et al. (2023) compared sedentary individuals with those that engaged in yoga or aerobics; they observed increased SVF performance in both exercise groups comparing to the sedentary control. Moreover, in a longitudinal study, Aichberger et al. (2010) followed individuals aged 50+ for 2.5 years and found that those who were physically inactive experienced a greater decline in cognitive abilities and SVF test performance compared to those engaging in either high or moderate levels of physical activity. Studies conducted on patients with various neurocognitive disorders have also demonstrated significant improvements in individuals' SVF performance when exercise programs were integrated into their daily routines; this was seen in Parkinson's disease (Cruise et al., 2011), in Alzheimer's disease (Öhman et al., 2016), and in mild cognitive impairment (Baker et al., 2010).

Leisure Activities: There are many forms of intellectually based leisure activities that are pursued for personal enrichment, knowledge acquisition, and mental stimulation. They include reading, puzzle solving (crossword puzzles, Sudoku, logic puzzles), writing, art-making (painting, drawing), engaging with music (playing an instrument, listening to music), and playing strategy games (Chess, Go). Numerous studies have found that these activities, with which people fill their free time and spend enjoyable time, can also have positive effects on cognitive abilities. Iizuka et al. (2019) conducted a review study on the effects of leisure activities that suggested that improvements in many cognitive domains were observed in 13 out of 20 studies. Kochhann et al. (2018) investigated *reading and writing* habits in healthy participants and patients with mild cognitive impairment and Alzheimer's disease, highlighting the positive effect that such activities have on verbal fluency tasks. In their 8-week follow-up study, Young et al. (2015) divided dementia patients into two groups, *art-viewing* and *art-making*, and observed a significant reduction in disfluency in both groups, but especially in the art-making participants. Yan et al. (2021) confirm the positive effects of *art therapies* for SVF in patients with mild cognitive impairment. In a longitudinal study with a large group of elderly European individuals, Cegolon and Jenkins (2022) highlighted the positive effects of different types of stimulating activities on SVF: (a) reading books,

magazines or newspapers, (b) completing crossword or Sudoku puzzles, (c) playing games (e.g. chess and cards). Grabbe (2011) also highlighted a positive correlation between SVF performance and solving Sudoku puzzles. Furthermore, playing board and card games were associated with better SVF performance in the study conducted by Estrada-Plana et al. (2021). As a result, studies on different intellectual leisure activities have shown that such activities have positive effects on executive function, especially SVF.

Chapter 3

Computational Linguistic Analysis of Semantic Verbal Fluency Data

3.1 Overview

Whereas the previous chapter discussed manual analysis methods, this chapter will focus on computational linguistic approaches to fine-grained analysis of SVF sequences. First, we give an overview of the algorithms and methods that have been used in the literature. Next, we describe the approach implemented in our Colombian Spanish SVF analysis study (see Chapter 5) and Turkish SVF analysis study (see Chapter 7). Our approach is based on vector space modelling, which is widely used in computational linguistics and computational semantics. In the chapters describing our studies, we will use a simple bigram-based method as a baseline for assessing the performance of the vector space modelling method, which is more resource-intensive.

3.2 Word Frequency Method: Bigram

Fürnkranz (1998) defines n-grams as a technique that examines consecutive words in a set-of-words approach. In this approach, the corpus is represented by a group of words, where n can take the values 1, 2, 3 (and so on); in these cases, it is called unigram, bigram, or trigram, respectively. Wolters et al. (2015) proposed a very simple baseline method based on counting the frequency of **bigrams** in SVF sequences. The idea is that this allows us to capture typical word association patterns for this particular task without the need to analyse large existing language corpora. In this method, there is no need to train models, so it does not require high CPU (central processing unit) power.

The corpus to be used with this technique comprises sequences of names produced by participants which vary depending on categories; animals, vegetables, and so on.

In the bigram method, the researcher counts how often each consecutive pair of words (**bigrams**), occurs in the dataset of SVF sequences obtained from all participants. The idea here is that an increase in the frequency of a bigram implies that individuals share a similar relationships in their word representations (Wolters et al., 2015). If a bigram occurs frequently (e.g. 'cat-dog'), the words are likely to share a semantic or pragmatic link. If a bigram occurs very rarely, such as 'cat-scorpion', this may indicate that the second word in the sequence is the start of a new cluster, or it may indicate a semantic link between the two words that is specific to that speaker.

The selection of 1 as a threshold is motivated by Zipf's Law (Zipf, 1936). For text, Zipf's law implies that when a set of words is listed in decreasing order of frequency, the rank of a word in this list is inversely proportional to its frequency, following an inverse power law distribution. Thus, words with no particular semantic links (i.e. words at switch locations) are likely to co-occur with a frequency of 1.

For relatively small SVF corpora that consist of tens or hundreds of sequences, a bigram frequency of 1 has proved to be a good approximation of switch locations. It is not clear whether this is also the case for substantially larger datasets containing thousands of SVF sequences. Due to the limited vocabulary used for the sequences, the number of word pairs with frequency 1 should decrease substantially.

The steps of the method are explained as follows:

1. Each consecutive word pair (bigram; e.g. dog-cat or cat-rabbit) is determined from the SVF sequences produced by participants.

2. The number of times each word pair occurs in the SVF data produced by all participants is calculated. The lowest frequency is equal to 1; this value indicates that the word pair was observed only once.

3. The last step is creating clusters and switches based on the bigram frequency scores. If the frequency score equals 1, a switch is created. All scores higher than 1 indicate that the words belong to the same cluster.

It should be noted that the order is significant in calculating the frequency score. That is, cat-dog and dog-cat are not the same. Hypothetically speaking, if the frequency score of dog-cat is 95 and cat-dog is 78, the word 'cat' was produced 95 times after the word 'dog', and participants produced the word 'dog' after the word 'cat' 78

times. Figure 3.1 provides an illustration of the bigram method for one SVF response produced by a participant, showing how clusters are formed using bigram frequency scores in three steps: (1) Participant response, which is the raw SVF sequence produced by the participant, (2) finding bigrams and frequencies by searching for animal pairs in the SVF dataset, and (3) creating clusters and switches based on the frequency score.

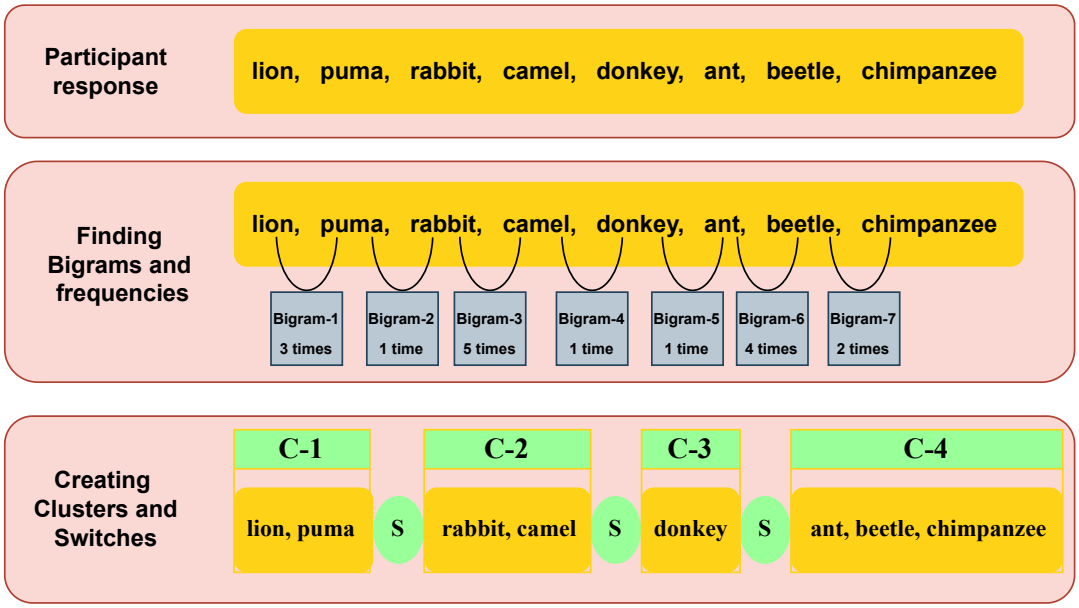


Figure 3.1: Steps of the bigram method: creating clusters and switches using bigram frequency scores of adjacent animal names. C: cluster, S: switch

In the bigram method, the frequency scores depend on the size of the dataset. Studies have shown that word production is affected positively by education (Ratcliff et al., 1998; Capitani et al., 1999) but negatively by aging (Burke and Shafto, 2004; Mortensen et al., 2006; Kempler et al., 1998). Moreover, as the level of education rises, the frequency of switching increases (Troyer, 2000). Younger people create more switches than older people (Troyer et al., 1997). Therefore, increasing the number of participants or having a young, well educated sample will increase the number of words produced (c.f. Section 2.5.1), and this will affect the frequency of word pairs. For example, having participants that generate many uncommon or infrequent words will result in word pairs that are only observed once in the entire dataset thus increasing the number of switches and creating more clusters. Briefly, datasets with a large variety of words can skew the results, so the bigram method can yield more successful results in datasets with a more balanced demographic distribution.

The bigram method will be applied to two different SVF datasets and will be examined in the respective chapters: (1) the Colombian Spanish language dataset in Chapter 5 and (2) the Turkish language dataset in Chapter 7.

3.3 Lexical Database Method: WordNet

Previously, in order to explore the similarities between given words, computational systems were created based on detailed lexical semantic databases, the best known example of which is WordNet (Miller, 1995). The original version of WordNet was created for English; detailed lexical maps are also provided for some other languages, including Arabic, Chinese, and Danish, but not all languages are represented. A full list of WordNets in the world can be found at the Global WordNet Organization webpage [1].

In these dictionaries, different word readings are represented with definitions including part-of-speech (POS) tags such as noun (n), verb (v), and adjective (adj). Figure 3.2 lists synonyms for the word 'mouse' using English WordNet[2]. 'Mouse' can denote an animal or a computer tool, depending on the given sentence, therefore predefined rules are required to capture the desired meaning.

Apart from serving as a thesaurus, WordNet contains hierarchical structures of semantically related words grouped using synsets (sets of cognitive synonyms). The networks between words in WordNet were created based on a super-subordinate hierarchical relation called ISA. ISA (is-a) relation focuses on lexical taxonomies and reflects the relation of two words regarding hyponymy (Liang et al., 2017). For instance, 'mouse' is-a 'rodent', and rodent is an upper class. However, each upper class can include own subsets; 'field', 'house', and 'harvest' exemplify the 'mouse' node. The hierarchy relationship is visualised in Figure 3.3, which also shows how the shortest path between two nodes is calculated.

Papers relying on WordNet use the shortest path in the WordNet graph to calculate word similarity score between words using Wu and Palmer similarity algorithm. In order to establish clusters, a threshold value needs to be determined to split words as a switch location, as in the bigram method (threshold value was 1 in bigram). For instance, Pakhomov et al. (2012) conducted a WordNet-based analysis on SVF se-

[1]The Global WordNet Organization webpage `http://globalwordnet.org/resources/wordnets-in-the-world/`

[2]WordNet output captured from `http://wordnetweb.princeton.edu/perl/webwn`

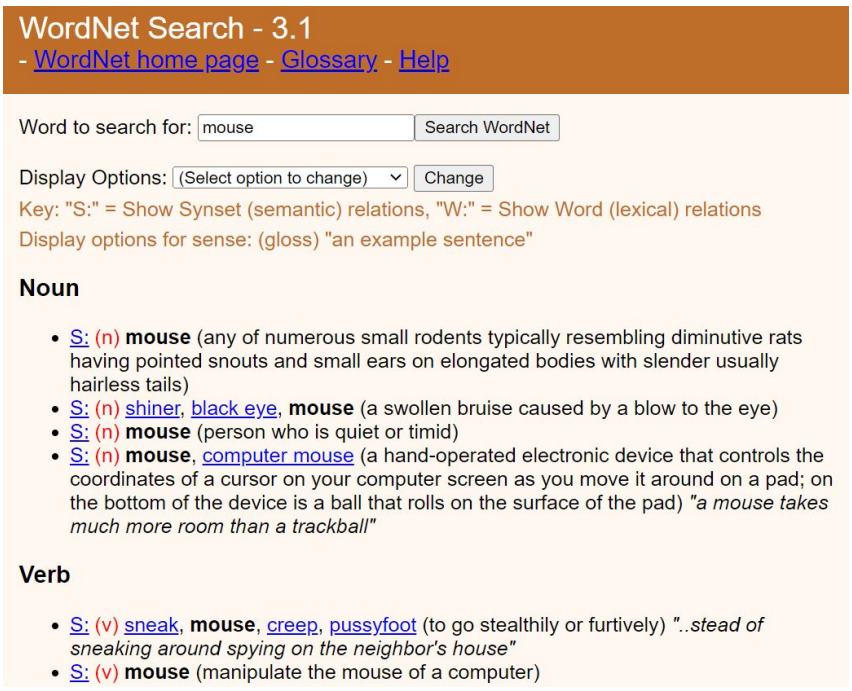


Figure 3.2: WordNet glossary representations for the word ‘mouse’, captured from online version of WordNet (Fellbaum, 2010).

quences collected from MCI and probable AD (high AD severity) groups, revealing a significant difference between two groups in terms of clustering and switching features. In another study, Paula et al. (2018) compared groups with various levels of cognitive impairment to healthy participants (binary classification: Control vs. Mild Cognitive Deficit). They observed that the features obtained through WordNet were successful in distinguishing the groups, although their comparisons showed that the word embeddings approach (via Glove and PMI) described in the next section (see Section 3.4) achieved higher scores in some cases.

In order to overcome the language barrier in WordNet, Paula et al. (2018) translated the SVF data collected in the original language into English. However, this approach lacks an established standard for local animals or words that do not have direct equivalents in English. Furthermore, it should be noted that a predefined hierarchical structure with predefined meanings implies that the relationship between two words will always have the same distance. This is a concern because words may not have the same contextual meaning in every language, meaning that the relationship between, for instance, ‘pig’ and ‘dog’, may differ crosslinguistically. For example, ‘pig’ is a part of the set of farm animals in some countries, whereas a Turkish person might consider it a wild animal. Therefore, if the collected data is translated into English directly and the

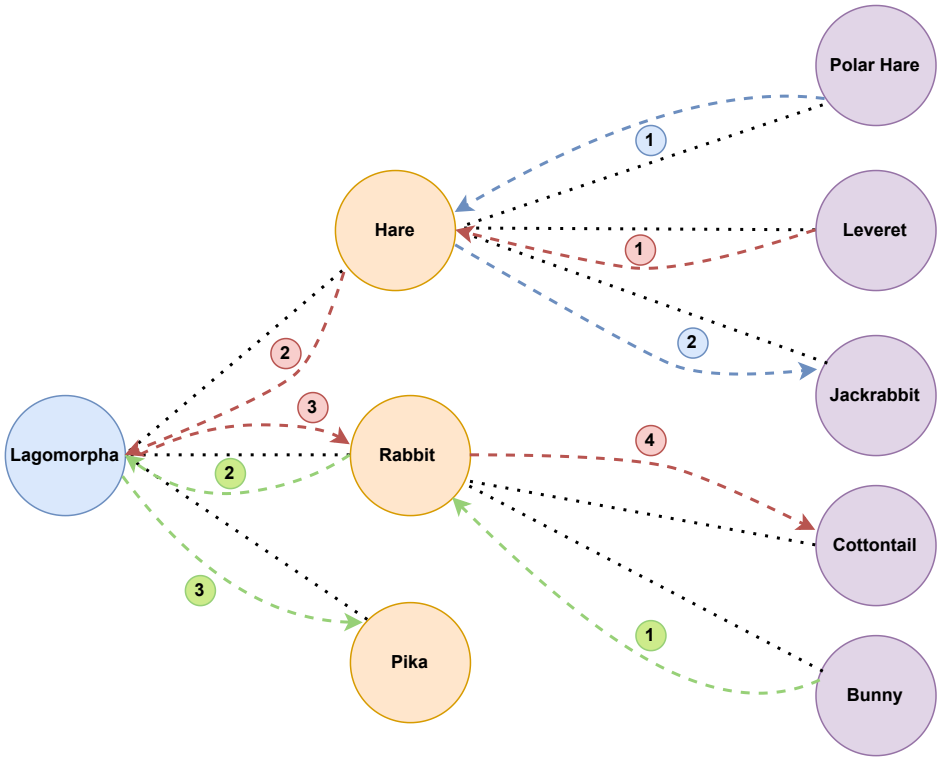


Figure 3.3: A WordNet hierarchy shows the shortest path between animal names (adapted from (Pakhomov et al., 2012)). The red line distance equals 4 for leveret-cottontail, the blue line distance is 2 for polar hare-jackrabbit,and the green line distance is 3 for bunny-pika.

similarity relationships are generated based on English, the results may not be entirely accurate.

Another limitation of WordNet is that it requires the researcher to manually choose the correct sense of a word from synsets in order to establish the relationships between words. For instance, when dealing with the word 'mouse', the researcher needs to decide whether it represents an animal or a computer tool; this can be labour-intensive, especially for extensive datasets.

3.4 Vector Space Modelling

Another way to describe the semantics of a word is with **word embeddings**, which are numerical vectors generated from a given set of texts (a corpus) so that words which are similar in meaning are also close in the vector space (Arora et al., 2016; Khattak et al., 2019). Word embeddings are sensitive to the length and coverage of the

corpus. Therefore, if a word is frequently employed across various contexts in a corpus, the resulting word vector is more likely to represent the meaning of the word better compared to a corpus containing only a few rare examples of the word. Harris (1954) hypothesised that if words are organised based on well-defined features, it is possible to create group of words that are close to each other in terms of meaning. Based on Harris' hypothesis, the vector space model (VSM) approach infers the relationship of words from the context in which they are used. In VSM, word embeddings are represented through a set of dots in a high-dimensional space; two words tend to have a similar meaning if they are close to each other (co-occurrence) (Erk, 2012). Therefore, the main aim is to create better represented vector space models.

Word representation techniques are the most important part of VSM. The closer the vector is to representing the meaning of the word, the better the obtained results will be. There are plenty of ways to calculate embeddings, which have emerged and evolved over time and range from frequency to meaning. They have been investigated by Turney and Pantel (2010) extensively. We will describe the word representation techniques hierarchically using Turney and Pantel (2010) as a guide, then we will consider developments that have emerged lately and examine common examples of VSM usage in the field. We will break up the discussion and categorise word embedding techniques as follows:

1. Count-based (non-semantic) word embeddings
2. Static (semantic) word embeddings
3. Contextual (semantic) word embeddings

Since this thesis primarily focuses on Word2vec, we will first provide a general overview of all common word embedding algorithms, including their architecture, then subsequently delve into an in-depth examination of Word2vec.

3.4.1 Count-Based (Non-Semantic) Word Embeddings

The simplest way to create a vector for a word or phrase is to look at the statistics and frequency of co-occurrence with nearby words or neighbours in the given corpus (Arora et al., 2016). The main drawbacks of this method is that it lacks information on the semantic relationships between words. Common techniques are described and exemplified below:

- **Bag of Words (BOW):** This is the basis for representing texts in terms of word frequency, without focusing on order; this is also called **Term Frequency (TF)** or **Term-Document Matrix**, which was developed by Salton et al. (1975). In this model, the entire document or specific sections thereof can be labelled as 'bags', and the focus is on how many times a word occurs in each bag, without normalisation of word derivations (Qader et al., 2019). Therefore, 'begin' and 'beginning' are considered different words. The term frequency formula is given below (Equation 3.1), where t is a term in a document d and w refers to words in the document.

$$tf(t,d) = \frac{f_d(t)}{\max_{(w \in d)} f_d(w)} \tag{3.1}$$

  Topic modelling algorithms were designed based on a BOW technique due to its basic representation and simplicity: **Latent Semantic Analysis (LSA)** and **Probabilistic Latent Semantic Analysis (PLSA)** (Wu et al., 2010). Topic modelling aims to classify documents using an unsupervised approach based on the frequency of words used together in given texts while discovering their theme (Kherwa and Bansal, 2019).

  There are some studies that have investigated clustering behaviour in SVF sequences using LSA approach. Prud'hommeaux et al. (2017) compared children with Autism Spectrum Disorder versus those with typical patterns of development (controls). They found significant differences in automatically derived cluster size, but not in manually derived ones between two groups. They suggested that LSA gives better insights for differentiating groups compared to the manual Troyer method. In their 20-year longitudinal study on individuals who were healthy at the beginning of the study, Pakhomov and Hemmy (2014) found that cognitive decline with age was observed more clearly in LSA-based clustering approaches using cluster size compared to the traditional word count approach.

- **Term Frequency-Inverse Document Frequency (TF-IDF):** As we discussed previously for BOW, the term frequency (TF) value is given by the frequency of a word in a specific document (which can be a paragraph, article, etc.); the higher the frequency of a word in a document, the higher the TF value (Qaiser and Ali, 2018). However, if a term is rare across the entire corpus (collection of docu-

ments), the **document frequency (DF)** value is low but its **inverse document frequency (IDF)** is high (Turney and Pantel, 2010). An IDF-weighting function was initially proposed by Sparck Jones (1972); it is given in Equation 3.2, where D represents the corpus, d donates documents, and t is a term. Later, various functions were added by Salton and Buckley (1988) and Singhal et al. (1996). The product of TF×IDF (see Equation 3.3) gives us the real importance of the word, called TF-IDF (also known as **document-term matrix**). This is based on Shannon (1948)'s hypothesis that the impact of rare events is greater than that of expected events (via Turney and Pantel (2010)). The primary objective is to reduce the influence of commonly observed stopwords such as 'the', 'a', 'an', and 'of', and highlight the rare words that carry more meaning in documents.

$$idf(t,D) = ln\left(\frac{\mid D \mid}{\mid \{d \in D : t \in d \mid}\right) \tag{3.2}$$

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \tag{3.3}$$

- **Pointwise Mutual Information (PMI):** PMI calculates the likelihood of two different terms (x and y) appearing together (joint probability) in a document based on the probabilities of each term appearing independently, which was proposed by Church and Hanks (1990). PMI focuses more on co-occurrence statistics, while Tf-IDF centres on a given term. The formula for PMI is given below in Equation 3.4, where x and y represent two words. The common point between PMI and Tf-IDF is that they reduce the bias caused by frequently observed values in a dataset.

$$PMI(x,y) \equiv log_2 \frac{P(x,y)}{P(x)P(y)} \tag{3.4}$$

  Paula et al. (2018) used PMI to distinguish patients with different levels of cognitive decline from controls and found that PMI outperformed WordNet based on a clustering feature. PMI also significantly outperformed the Troyer method (manually-derived) by a margin of 11 points. The results indicate that PMI is a successful method for identifying differences in grouping strategies (clusters) between patients and healthy individuals.

The common thread among these techniques is that they represent a word with an unweighted (BOW) or weighted (Tf-IDF, PMI) vector that reflects their frequency in the

given documents, regardless of meaning relationships. In this respect, these methods are non-semantic. In computational linguistics, count-based word embedding techniques are considered fundamental as they are common and easy to adapt and upgrade. However, they have not been widely used in the field of SVF, except for the limited number of examples shared above.

3.4.2 Static (Semantic) Word Embeddings

Methods within this group, which can also be referred to as **Classic Word Embeddings**, offer an algorithmic structure compared to count-based methods. Static word embeddings have introduced pre-trained models, which generate vectors for words from unlabelled data via statistical methods (Mikolov et al., 2017). Pre-trained models include a downloadable collection of saved static vectors which was created by someone. Also, it is allowed to be trained by researchers using relatively large corpora of text data (such as Wikipedia articles), but the output consists of fixed vectors for each word (Neelima and Mehrotra, 2023; Gupta and Jaggi, 2021). Using different corpora can assure different word vectors for words (Pham and Le, 2018). Although these methods have different algorithmic structures, their similar representational success leads to adaptation to various fields, such as cognitive science (Naseem et al., 2021). Below we will describe the most common algorithms and their representation approaches as well as the examples for SVF field.

Word2vec: Word2vec is the first statistical word embedding were proposed by Mikolov et al. (2013a,b). Word2vec method uses a feed-forward Neural Net Language Model (NNLM) (Bengio et al., 2003) to train words through untagged and unlabelled data. This is the well-known state-of-the-art model to integrate neural network approaches into the word embedding. The Word2vec team has published pre-trained model, described in the original article, Mikolov et al. (2013a) as a resource for other researchers and provided details about the model in google code archive [3].

There are two types of architectures in Word2vec: (1) CBOW, which stands for Continuous Bag-of-Words Model, and (2) Skip-gram, which stands for Continuous Skip-gram Model. CBOW was the first proposed model, and it is faster one. It takes input as a group of words based on a Bag-of-Words (BOW) approach, which means that no information about the order of words is preserved Mikolov et al. (2013a). In Skip-

[3]Word2vec pre-trained model: `https://code.google.com/archive/p/word2vec/`

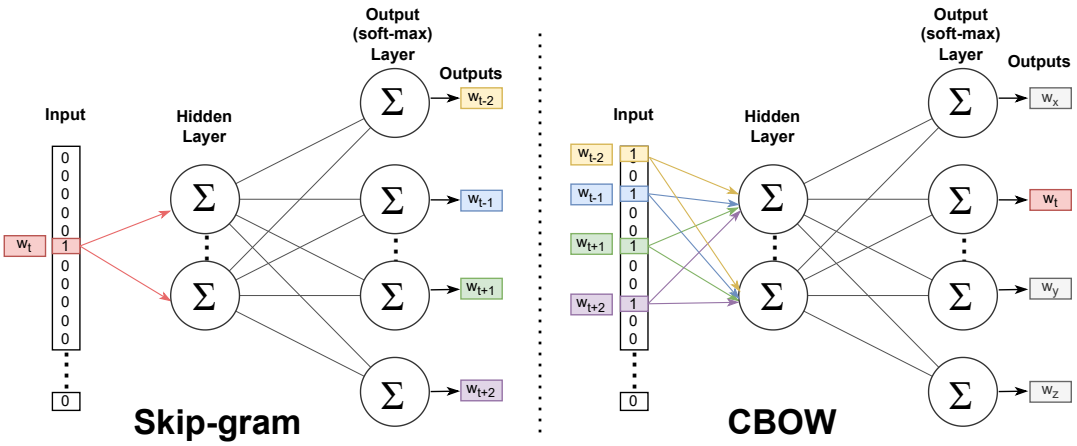


Figure 3.4: Visual representation of Word2vec architecture: CBOW and Skip-gram. The figure was adapted based on a main article by Mikolov et al. (2013a) and inspired by Rong (2014) and Tshitoyan et al. (2019).

gram, a single word is used as input to predict a group of nearby words which makes it computationally intensive due to the need for making multiple predictions Mikolov et al. (2013a). Figure 3.4 simplifies the network structure of both models. Both methods use the same linear neural network and softmax function. In Equation 3.5, you can see the main Skip-gram formulation of Word2vec, taken from Mikolov et al. (2013b), which takes word as one hot encoded vector: only input word w_t codded with 'one' and all other neighbour words codded with 'zero'. The output gives the probability of a word $w_{t+j}|w_t$ occurring given a previous word w_t using softmax function. For example, w_t is the word 'dog' and w_{t+j} is the word 'cat'. In this case, the $w_{t+j}|w_t$ expression indicates the probability of 'cat' coming after 'dog'.

$$p(wo|w_I) = \frac{\exp(v'_{w_o} T v_{wI})}{\sum w = 1^W \exp(v' w_o T v_{w_I})} \tag{3.5}$$

Word2vec has gained attention in the SVF field due to its success in capturing group differences in various research studies. One of the early examples of studies that investigated the Word2vec model and created a state-of-art analysis on French SVF dataset was Linz et al. (2017a). In order to compare patients with mild cognitive impairment to healthy participants, they trained a classifier using clustering and switching features derived from both the Word2vec automated method and the Troyer baseline method. Word2vec outperformed the baseline and distinguished groups clearly. In another study of Linz et al. (2017b) used Word2vec as an additional feature-set generator, using clustering and switching features to train their regression models among

other linguistic and vocal features to estimate patients' scores on the Mini Mental State Examination (MMSE) and standard Dementia Rating (CDR). Moreover, König et al. (2018a) compared how Word2vec outputs correlated with manual Troyer results in terms of derived cluster- and switch-based features. The findings showed a high correlation (r = 0.9) between Word2vec and Troyer's method in terms of extracted features, which were the numbers of clusters and switches. Word2vec is as successful as the Troyer method in distinguishing between dementia patients and healthy people. Similarly, Prud'hommeaux et al. (2017) compared Word2vec and Troyer with clustering to distinguish children with autism spectrum disorder and typical development. Troyer features found no differences between both groups, but Word2vec indicated significant differences, especially for cluster length. These findings demonstrate that the Word2vec method outperforms Troyer's approach in the clustering–switching technique, and that heuristic threshold values for creating clusters can yield better results compared to Troyer's limited predefined animal groups.

Studies have typically trained their own Word2vec models based on the language in which their SVF data were collected and reported findings from a single combination of hyperparameters (Linz et al., 2017a,b; König et al., 2018a; Prud'hommeaux et al., 2017). In contrast, Kim et al. (2019) explicitly discussed the model performance with different combinations of hyperparameters. Kim et al. (2019) tested models on a dataset obtained from healthy participants using two categories, animals and fruits, and two languages, English and Korean. For Italian, Di San Pietro et al. (2021) and Di San Pietro et al. (2023) found that the features produced by the Word2vec model positively affect the success of the classifier to separate two groups of participants. Other studies have used pre-trained word vectors created by Mikolov et al. (2013a). In English, for example, Mikolov's vectors have been used to investigate clustering and switching in schizophrenia (Lundin et al., 2020) and early-stage psychosis (Lundin et al., 2022). In another study, Swiss-German words in the SVF data were translated into English so that a pre-trained Word2vec model could be used (Saranpää et al., 2022). In this example, the researchers compared healthy individuals versus patients with early Alzheimer's disease and amnestic mild cognitive impairment (AMCI), by exploring clustering-switching features; they demonstrated that the AD group tended to switch less than patients with AMCI and healthy people. They also investigated how many times participants returned sub-groups of animals (e.g. birds, pets, fishes) that they had visited before. In that case, AMCI and healthy controls revisited sub-groups more than AD patients did.

As Word2vec is the main computational approach investigated in this thesis, it will be thoroughly discussed in Section 3.5, including a description of our pipeline.

GloVe: GloVe was introduced by Pennington et al. (2014) shortly after Word2vec as another word embedding model based on global word co-occurrence counts. The name 'GloVe' stands for 'global vectors' to emphasising the fact that the model learn from all examples of the word in a given corpus. Like Word2vec, the GloVe team has made available different pre-trained models containing various tokens and word counts to researchers in their website [4]. While GloVe is based on the same direct prediction (predicts one word from another) method as Word2vec, Word2vec utilises local statistics around the target word, whereas GloVe relies on global statistics by combining co-occurrence information from different examples to create a larger matrix (Dharma et al., 2022). GloVe consists of two steps first is creating global matrix and second is factorisation which lowering the dimension of matrix then each row represents a word (Naseem et al., 2021). The GloVe model function is given in Equation 3.6, where V=vocabulary size, X=global matrix, X_{kj}=frequency of being together word k and j, X_k=total occurrences of work k, P_{kj}=probability of being together word k and j, w=a word embedding (input), $w^{'}$=the context words (co-occurrences).

$$J = \sum_{k,j=1}^{V} f(X_{kj})(w_k^T w_j^{'} + b_k + b_j - \log X_{kj}) \tag{3.6}$$

Pennington et al. (2014) states that GloVe outperformed Word2vec using the same corpus, word count, and training time on two tasks: similarity and named entity recognition. On the other hand, Levy et al. (2015) and Hossain et al. (2021) emphasised that the training time and memory consumption of GloVe is massive rather than Word2vec. Both models are considered state-of-the-art, but in the SVF field, GloVe is relatively less experimented than Word2vec.

Paula et al. (2018) compared individuals with various levels of cognitive impairment with healthy controls, deriving different features to be used as input for Random Forest classifiers on Brazilian Portuguese. They created their own GloVe embedding models with a window size of 7 and a vector dimension of 300. Notably, their research showed that features based on GloVe similarity-based clustering outperformed those based on WordNet and PMI when it came to distinguishing between individuals with Multi-domain Mild Cognitive Deficit and healthy controls.

[4]GloVe pre-trained models: `https://nlp.stanford.edu/projects/glove/`

In a separate study conducted by Pietrowicz et al. (2019), GloVe embeddings were also employed. However, the focus here was on identifying healthy individuals who were at risk of developing schizophrenia. To achieve this, various classifiers were trained with features derived from GloVe-based similarity calculation between words, and the study successfully identified individuals with an 11% or higher likelihood of developing the disease.

FastText: The main disadvantage of the predecessors of the field, GloVe and Word2vec, is that if a word does not exist in the corpus, a vector output cannot be obtained for that word (Naseem et al., 2021). To address this difficulty, Bojanowski et al. (2017) proposed a new model named FastText which is based on a CBOW baseline of Word2vec. The improvement of FastText lies in breaking words into sub-word components which is called n-gram (sequence of characters) and representation of the word is sum of each n-gram vector (Dharma et al., 2022). In simpler words, words are divided into pieces regardless of their meaning, and each piece is expressed with a vector output. For example, 'elephant' can be separated 2-grams and the output sub-words are 'el', 'ep', 'ha', 'nt'. The vector of 'el' can be also used to achieve the vector of 'umbrella'. FastText is a good alternative for generating rare words from a morphological perspective, but it is weak in capturing semantic relationships (Choi and Lee, 2020).

In their study, Venekoski and Vankka (2017) evaluated three static word embedding techniques—Word2vec, GloVe, and FastText—in terms of semantic quality for Finnish, a low-resource language. Their findings revealed that Word2vec and FastText outperformed GloVe in many tasks, such as similarity judgment. Additionally, they compared various Finnish corpora as language sources including Wikipedia to train the models and found that mostly Word2vec performed better than GloVe and FastText in capturing conceptual relationships on different corpora.

In the field of SVF, some studies have applied the FastText method to discover word representations. Tröger et al. (2019) used FastText to capture semantic relationships between words and utilised a version of the Troyer method adapted for French (Troyer et al., 1997) for manual annotation. When examining clustering-switching based features, the study observed similar patterns between the Troyer method and FastText, indicating a high or moderate correlation. The findings showed that there were fewer switches in the AD group compared to the control and MCI groups. However, no group differences were identified in terms of mean cluster size. Both feature results are similar to the results obtained with the Troyer method. In another study on a French

population, Lindsay et al. (2021a) employed a combination of the features obtained with Troyer's method and those obtained using FastText. The probability of differentiating between MCI patients and healthy controls increased by 29% (measured by the area under the curve (AUC)) comparing to features only gathered from Troyer method. Other studies that have utilised clustering- and switching-based features obtained with FastText have similarly argued that these features enhance the decision-making process and provide valuable insights for clinicians in distinguishing between different groups (Garcia, 2023; Lindsay et al., 2021b).

3.4.3 Contextual (Semantic) Word Embeddings

Word2vec, GloVe, and FastText can successfully extract the semantic meanings between words from sequential lists of words (Dharma et al., 2022). The main drawback of these three methods is that they represent each word with a single vector, which is a limitation when dealing with homonyms such as 'bank', which can have different meanings depending on the context (Han et al., 2021).

Contextual word embeddings are word representations that create context-dependent dynamic vectors through optimisation or fine-tuning techniques on pre-trained vectors. The neural network-based initial examples, Context2Vec (Melamud et al., 2016), CoVe (90) (McCann et al., 2017), and ELMo (Peters et al., 1802) use the bi-directional long short-term memory (LSTM) to leverage the performance of Word2vec baseline and present a solution to capture context base meanings (Naseem et al., 2021). LSTM uses recurrent neural networks (RNN) and has been shown to be successful with sequential data, but it is hard to infer meaning from larger dependencies. However, transformer architecture-based methods such as GPT (Generative Pre-Trained Transformer) by OpenAI (Radford et al., 2018) and BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) emerged shortly after. Those methods are able to capture many possible semantic representation of words, but they need excessive amount of data and intense computing power, which makes them computationally expensive. Transformers can work with long sequences better than LSTM; their use of parallel processing makes them powerful. These models are often referred to as **large language models**, and comprehensive pre-trained models are readily available for use. Users can leverage these models by fine-tuning them with their own datasets, customising the existing models for their specific tasks.

Large language models have emerged as highly successful tools for uncovering

semantic meanings and exploring diverse word usages based on context. They are increasingly under investigation for various natural language processing (NLP) tasks, including information retrieval, summarisation, and language generation. However, in the specific context of the SVF field, these models have not yet been widely adopted. Research in this area remains limited, with only a few studies conducted, which are shared below.

In Alaçam et al. (2022), which used an SVF dataset with productions for ten categories, including animals, manual annotation clusters were created following the Troyer method and subgroups for other categories were manually created. Then, clusters were automatically generated using BERT, GloVe, ConceptNet, and FastText. The manual method was compared to the automatic methods using Pearson correlation, and some categories showed the highest correlation with manually curated clusters. The ConceptNet model appeared to be the most similar to manual methods, and Bert had the lowest scores in terms of determining the right switches. ConceptNet provides a semantic network for words and offers related words or words with semantic relationships. It is similar to WordNet but more comprehensive and not only related to lexical hierarchical meaning.

Nighojkar et al. (2022) focused on predicting which word would come after a given word in an SVF sequence rather than clustering-switching features. They employed eight different algorithm-based methods, including three Bert-based models and two static word embeddings (GloVe and Word2vec). Among these, RoBERTa emerged as the top-performing model, achieving a prediction accuracy of 86.4% for vegetables, and Word2vec and GloVe followed. MiniBert and DistilBert, on the other hand, exhibited relatively poor performance.

3.5 Our Approach: Word2vec

In this thesis, we use the Word2vec method proposed by Mikolov et al. (2013a,b), which is a well-known static word embedding technique that uses the VSM approach. The clustering-switching technique discussed in Section 2.4.2 depends on predefined groups and needs to be expanded, which makes it difficult to replicate. Therefore, Word2vec method is generally preferred today because it extracts the relationship between two words by looking at the sentences, paragraphs, and texts in which these words are used together. Some of the advantages of the Word2vec method are as follows.

- It depends on computer power rather than human labour and annotations.
- It can be replicated easily if the details of the system are described.
- It may require less computing power than contextual word embeddings.
- It captures up-to-date word meanings well through use of text collections that are updated regularly with newspapers, Wikipedia articles, transcripts of spoken language, and so on.

3.5.1 Building Word2vec Models

3.5.1.1 Wikipedia as a Corpus

Vector space models require a corpus of example sentences to create vectors for desired words. In this thesis, Wikipedia was used as an unlabelled data source for Word2vec word mapping. A Wikidump is a collection of up-to-date Wikipedia articles in a given language. Copies of the database are taken at different time periods and stored on servers provided by the Wikipedia Foundation. Wikipedia offers this service for all languages through volunteers [5].

3.5.1.2 Pre-Processing steps:

Data pre-processing is a crucial step for almost any type of data used to feed the algorithm, which can include images, text, and videos. The common view is that using appropriate pre-processing methods will eliminate data complexity and uncertainty. Here, we provide a general overview of the pre-processing steps for a Wikidump. The specific tools used for pre-processing will be explained in the data analysis sections in Chapter 5 for Spanish and in Chapter 7 for Turkish.

- **Data cleaning:** Wikidump data provides Wikipedia pages in a raw XML file format, which includes many tags and unnecessary elements. Since our target is to work on sentences in the XML files, the best is to extract the <title> and <text> tags first. Using Json format syntax gives us faster and easier execution than simply reading dump files, so we used json files to store plain text versions of articles. To extracting plain text out of a raw Wikipedia dump, we used scripts.segment_wiki provided by Gensim (Rehurek and Sojka, 2011).

[5] Wikipedia page describing how to download Wikidumps: `https://dumps.wikimedia.org/`

- **Lowercase:** With this technique, which is used as a standard in all text processing methods, the complexity of upper and lower case letters is eliminated. For example, 'dog' and 'DOG' are represented two different vectors, which increases the dimension of the space. After converting the text to lowercase and eliminating the different size of the letters, a single vector represents each word.

- **Punctuation:** Punctuation removal is another fundamental process for text analysis. 'dog' and 'dog!' will be assumed to be different words due to the punctuation. Removing punctuation ensures that these words are treated as the same and improves model results.

- **Stopwords:** Stopwords represent words that do not have a significant meaning in the sentence and do not cause a change in meaning when removed; they can be considered noise (Kaur and Buttar, 2018). They can be determiners, conjunctions, or prepositions, such as 'a', 'an', 'but', 'or', 'in', and 'on'. Stopwords vary by language. Removing stopwords reduces the complexity.

- **Stemming - Lemmatisation:** A word can appear in several different forms due to inflectional and derivational affixes. To decrease the complexity in the dataset, two methods are used to reduce words to the basic unit: lemmatisation and stemming. Stemming produces a common stem for all words with the same root, while lemmatisation simply removes prefixes and suffixes used for inflection and leaves the root of the word. Lemmas are dictionary words, but stems are not (Balakrishnan and Lloyd-Yemoh, 2014). Since Turkish is a morphologically agglutinative language, its words have many more affixes compared to English. Therefore, we investigated both lemmatisation and stemming with different libraries.

3.5.1.3 Model Hyperparameters

Word2vec proposes two architectures to predict word meanings: Continuous Bag-of-Words (CBOW) and Continuous Skip-gram (Skip-gram), respectively. Mikolov et al. (2013a) explains the difference between the two architectures as follows:

- **CBOW** predicts the desired word from the given set of words.

- **Skip-gram** predicts the related words from the given central word.

Toy text can be used to examine how CBOW and Skip-gram make predictions. Our toy text is 'Dog is a domesticated carnivorous mammal', the target word is 'domesticated', and the window size is 2. According to the bigram (2-gram) window size, the neighbours of the target word are 'is', 'a', 'carnivorous', and 'mammal'. Input is created in the form of the word pairs (domesticated, is), (domesticated, a), (domesticated, carnivorous), and (domesticated, mammal). In CBOW, the neural network trained with these inputs predicts the target word. However, in Skip-gram, neighbours need to be predicted from the target word, so the model takes the target word and tries to find nearby words. While doing this, the algorithm looks at the list of words that appear frequently with the target word and tries the possible words one by one until it reaches the neighbour words. CBOW is much faster than Skip-gram. Briefly, CBOW fills in the blank in 'Dog is a _____ carnivorous mammal', and Skip-gram fills in the blanks in 'Dog _____ _____ domesticated _____ _____'.

The input words of CBOW and the output words of Skip-gram are called **'window size'**. In other words, how many words are used as a group is determined by the n-gram method. If window size equals 5, it is called a five-gram. In addition to architecture and window size, the **dimension** should also be determined prior to vectoring to perform superficial or deep modelling in word predictions. While low dimensions make modelling simpler, denser dimensions reveal more detailed relationships. The selection of hyperparameters also directly affects training time and results (Lison and Kutuzov, 2017).

To capture the best model representation, Wikipedia was used as a corpus for training the models in our study, using different hyperparameter combinations taken from Kim et al. (2019). This resulted in 12 models from shallow to deep. Clusters and switches were created based on these models. Our parameters were:

- **Objective Function (Architecture):** 2 parameters; CBOW and Skip-Gram
- **Window-Size:** 2 parameters; w = 4, 10
- **Dimensions:** 3 parameters; d = 300, 600, 1000

Limitations in Hyperparameter Selection: The list of hyperparameters was replicated from Kim et al. (2019), in which the various parameters were previously tested for English and Korean, leading to successful outcomes. Detailed hyperparameter selection requires computationally expensive deep optimisation techniques, which standard office computers may lack. In our study, operations were carried out with an

easily accessible computer with Intel Core i7-7500U CPU processor. We encountered that training time was significantly increased for models using the Skip-gram structure, especially those working with a window size of 10 and a dimension of 1000. Considering the dataset sizes, problems such as insufficient memory and the termination of model creation before completing the process are possible. Additionally, the Word2vec main paper offered a general-purpose model pre-trained on Google news releases with a dimension of 300 (Mikolov et al., 2013a), which is a superficial depth model. The fact that authors did not use very deep model training shows that simple models can also perform well. One should also consider that delving into deeper models may lead to potential issues, such as detaching the word from the overall context of the text and capturing rare meanings. For more comprehensive research, it is possible to train models with GPU-based computers, but a trade-off between time and expense should be adjusted.

In our study, words were highly unlikely to be used metaphorically. In this case, the use of models that do not delve too deep but rather strike a balance between deep and shallow are likely to be successful in capturing general meanings and testing models. For instance, considering the word 'snake', our goal is to find a vector that symbolises a real animal, so the figurative interpretation 'sly' is inappropriate for our study. However, both meanings can be true depending on the nature of the study.

3.5.2 Distance Between Word Embeddings: Cosine Similarity

Cosine similarity is the best-known measure for computing the distance between words. Kiela and Clark (2014) investigated vector space model parameters and components in their articles and stated that cosine is one of the best performing similarity metrics. In this study, cosine similarity was used to determine the relationship between two words. The cosine measure is calculated for every pair of consecutive animal names produced by one participant based on vectors generated by the Word2vec model. Equation 3.7 shows how to calculate cosine similarity metrics, where a and b are vector representations of two different words. The cosine similarity value ranges from -1 to +1. A value of -1 ($\theta = 180$) indicates opposite meanings, while +1 ($\theta = 0$) signifies that two words have the same meaning. A value of 0 ($\theta = 90$) indicates no relationship between the two words. Figure 3.5 shows how θ changes between two vectors and how cosine interprets meaning.

$$\cos\theta = \frac{\vec{a} \times \vec{b}}{\|a\| \times \|b\|} \tag{3.7}$$

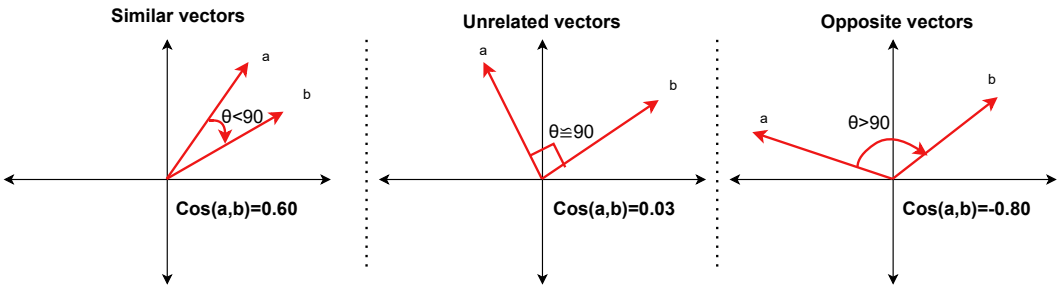


Figure 3.5: Cosine similarity angles between two vectors: *a* and *b*. The figure was adapted from Sarma et al. (2021)

According to the model's parameter selection, the similarity score for the same two words will vary in each model, even if trained with the same corpus (in our case, the same Wikipedia articles). For example, while one model may show a similarity of 0.90 for 'cat' and 'dog', another model with different hyperparameters may yield a similarity value of 0.70. Although the results may indicate different relationships, we expect to see similar patterns (such as similarity or dissimilarity). Furthermore, a higher similarity score for a pair of words indicates a close relationship between the words. Importantly, whether the pair is 'cat–dog' or 'dog–cat', it represents the same relationship. Moreover, it is important to consider the sampling capability of the corpus on which the model is trained. For instance, if there is a sufficient number of sample sentences for a specific word, the model will capture the meaning better. Word2vec models may fail to create a vector representation for rare words, and consequently, the cosine value may not be calculable.

3.5.3 Determining Cluster Boundaries: Threshold

Unlike the Troyer method, there are no predefined taxonomy groups available to determine clusters in automatic computational methods, so studies have determined threshold values for clusters based on the nature of their own datasets. Threshold values are used to split word sequences into clusters, which are determined heuristically based on a practical principle, according to Prud'hommeaux et al. (2017). Some studies have used different global threshold values calculated from the similarity scores obtained from the entire dataset for cluster boundaries: 0.5 and 0.6 in Farzanfar et al. (2018),

0.75 in Woods et al. (2016b); Rosenstein et al. (2015), 0.9 in Pakhomov and Hemmy (2014), and various different values experimented in Di San Pietro et al. (2023); Kim et al. (2019). However, Linz et al. (2017a) pointed out the difficulty of determining a global threshold value and instead chose to set individual (local) thresholds for each person. They used the average similarity score (0.5) across all words belonging to a person. In contrast with other studies relying on lexical features, Tröger et al. (2019) proposed a novel vocal threshold, which is the average waiting times between words.

Our studies in Colombian Spanish (Chapter 5) and Turkish (Chapter 7), examined twelve models with different combinations of hyperparameters; each model was tested on three global threshold values for the cluster boundary (the 25th , 50th, and 75th percentiles), creating a total of 36 cluster files. The aim was to determine which model produced results that were the most similar to the manually annotated clusters, using the Troyer method as the gold standard. As in the hyperparameter selection, Kim et al. (2019) was followed to determine threshold values as well.

A step-by-step breakdown of how SVF sequences produced by participants are clustered is given below, summarising the process so far. A hypothetically selected model will be used for illustration purposes: a 300-dimensional vector space, a window size of 4, and the CBOW architecture, using Wikipedia as the corpus (the model name is 300_4_CBOW briefly).

1. The model contains vector representations for words (if sufficient example exist) in a 300-dimensional space.
2. Word pairs said consecutively by each individual in the entire dataset are identified. For example, in Figure 3.6, possible word pairs include cat–dog, dog–rabbit, and rabbit–fish.
3. For each word pair, cosine similarities between the words are calculated using the vectors for the selected model. For example, in Figure 3.6, $Sim(cat, dog) = 0.678$ and the blue line shows all adjacent words' cosines for one participant.
4. The calculated cosine pairs are sorted from the highest to lowest similarity score.
5. Threshold values for the 25th, 50th, and 75th percentiles are determined from these sorted similarity scores. In Figure 3.6, the red line represents the 50th threshold value, which is 0.394 for the example.
6. Clusters are created for each person based on the threshold value. If the similarity between two words exceeds the threshold, they belong to the same cluster. Otherwise, a switch is placed between word pairs. Figure 3.6 illustrates how

cosine scores are used to create clusters, and switches are shown in Figure 3.7.

7. For each participant, three cluster-switch outputs, one for each threshold (the 25th,50th and 75th quartiles), are created based on each model (300_4_CBOW in here).
8. This process is applied for all models (12 models in total).

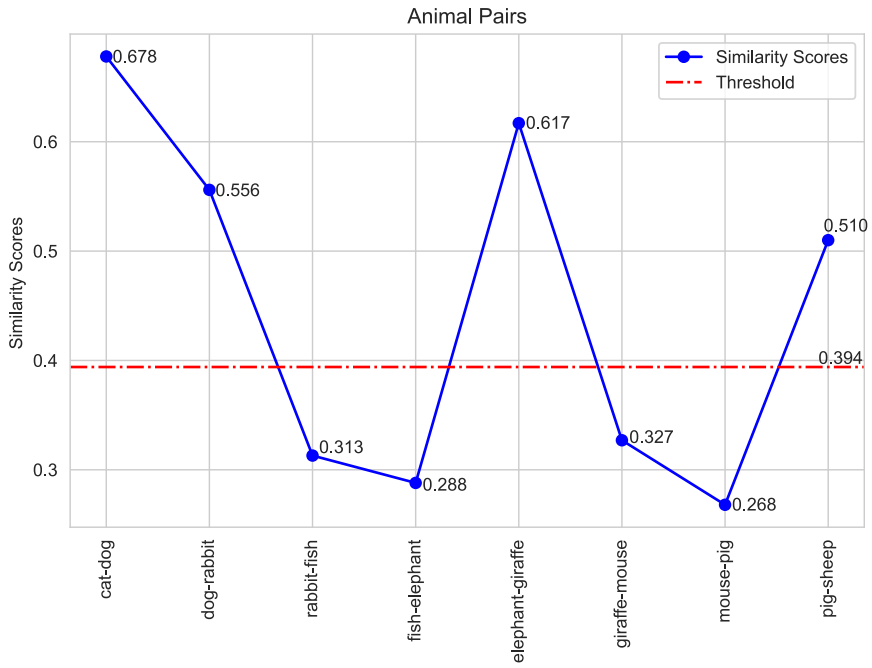


Figure 3.6: Example of Cosine similarity measures for every consecutive words in a sequence and how threshold value cuts the values.

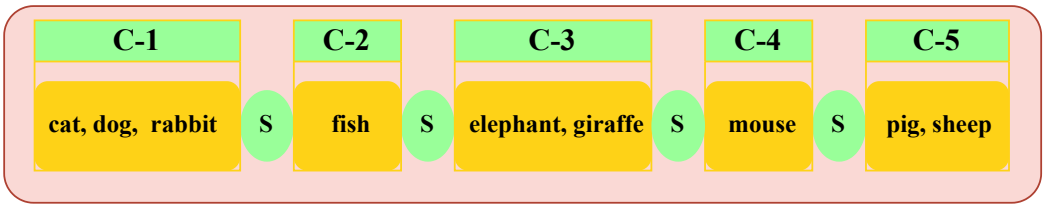


Figure 3.7: Clusters and switches created according to the threshold score determined from the vector space model. C:cluster, S:switch

Chapter 4

Semantic Verbal Fluency in Native Speakers of Turkish: A Systematic Review of Category Use, Scoring Metrics and Normative Data in Healthy Individuals

Note: The following chapter has been submitted to the Journal of Clinical and Experimental Neuropsychology (CEN-RA 23-107). We were invited to submit a revised version on August 5, 2023. The text reflects the revised version of the paper, which will be submitted after submission of the thesis. The PROSPERO registration of the protocol for the systematic review can be found in Appendix A.2, and an early version of the work was published at the Conference of the Psychonomic Society (c.f. Appendix A.1).

Author Contributions: The reviewers jointly designed research questions, search strategy, database selection and data extraction template. RYK conducted the database search and uploaded articles into Covidence. Conflicts were resolved through team discussions. All studies were screened by two team members: RYK and one team member, MW or SM for English articles, KK for Turkish articles. Data extraction was carried out by RYK with support from KK, and MW and SM were consulted for any queries. First draft of the study was created by RYK and edited by MW and SM.

4.1 Abstract

Introduction: Semantic verbal fluency (SVF) is a widely used measure of frontal executive function and access to semantic memory. SVF scoring metrics include the number of unique words generated, perseverations, intrusions, semantic cluster size and switching between clusters, and scores vary depending on the language the test is administered in. In this paper, we review the existing normative data for Turkish, the main metrics used for scoring SVF data in Turkish, and the most frequently used categories.

Method: We conducted a systematic review of peer-reviewed papers using Medline, EMBASE, PsycInfo, Web of Science, and two Turkish databases, TR-Dizin and Yok-Tezthat. Included papers contained data on the SVF performance of healthy adult native speakers of Turkish, and reported the categories used. Versions of the SVF that required participants to alternate categories were excluded. We extracted and tabulated demographics, descriptions of groups, metrics used, categories used, and sources of normative data. No quality assessment was performed.

Results: 1400 studies were retrieved. After deduplication, abstract, full text screening, and merging of theses with their published versions, 121 studies were included. 114 studies used the semantic category "animal", followed by first names (N=14, 12%). All studies reported word count. More complex measures were rare (perseverations: N=12, 10%, clustering and switching: N=5, 4%). Four of seven normative studies reported only word count, two also measured perseverations, and one shared category violation and perseverations . Two normative studies were published in English.

Conclusions: There is a lack of normative Turkish SVF data with more complex metrics, such as clustering and switching, and a lack of normative data published in English. Given the size of the Turkish diaspora, normative SVF data should include monolingual and bilingual speakers. Limitations include a restriction to key English and Turkish databases.

Prospero registration reference: CRD4202020158, last search: June 14, 2022. Funder: Ministry of National Education of Türkiye

4.2 Introduction

Semantic verbal fluency (SVF) is a widely used neuropsychological assessment to examine frontal executive functions (EF) and access to semantic memory (Patterson et al., 2011). For SVF, participants are asked to produce as many words of a given category as possible in a short amount of time, usually 60 seconds. The most frequently used category is animals, but other categories such as supermarket items or fruit and vegetables have also been used. SVF is often paired with another type of verbal fluency test, phonemic or letter fluency. In letter fluency, the participant is asked to produce as many words as possible that start with a specific letter. In English, these letters are usually F, A, and S or C, F, and L (Amunts et al., 2020).

SVF is part of several standard cognitive assessment batteries, such as the Addenbrooke's Cognitive Examination (Mathuranath et al., 2000b) or Delis-Kaplan Executive Function System (D-KEFS) (Delis et al., 2001a). Many conditions such as Alzheimer's Disease (Monsch et al., 2020; Gomez and White, 2006), Parkinson's Disease (Piatt et al., 2010; Raskin et al., 1992b) and traumatic brain injury (Raskin and Rearick, 1996b; Zakzanis et al., 2011; Kavé et al., 2011) consistently affect participants' SVF performance. Fluency tasks are commonly used to detect frontal executive dysfunction with frontal patients performing more poorly than posterior patients and healthy controls (Milner, 1964; Baldo and Shimamura, 1998; Stuss et al., 1998; Troyer et al., 1998a). However, some SVF studies have reported impairments associated with both frontal and posterior lesions (Robinson et al., 2012; Vilkki and Holst, 1994; Stuss et al., 1998).The meta-analysis by Henry and Crawford (2004) demonstrated frontal patients were impaired on both phonemic and semantic fluency but temporal patients showed a larger semantic than phonemic fluency impairment.

The most commonly used scoring metric for SVF is the total number of correct unique words produced within the time-limit (Ardila et al., 2006). However, SVF is considered a multifactorial task and simply considering the total number of correct words is not thought to be sufficient to fully capture a participant's performance. Some researchers also report the number of words produced in the first, second, third, and fourth quarter of the 60 second time-frame. Errors can provide a qualitative measure of SVF performance, and are divided into two subtypes: intrusions and perseverations (Raboutet et al., 2010). Intrusions or category violations are additional words that do not belong to the desired category (Raskin and Rearick, 1996a). Perseveration refers to the repetition of words and is divided into three types (Sandson and Albert,

1987; Pekkala et al., 2008): recurrent (repeated words separated by other words, e.g., cat, dog, fox, cat), stuck-in-set (repetition of words from a previous category, such as animals, in the current category, such as supermarket items), and continuous (repeated production of a single word, e.g., cat, cat, cat). Generally, all repeated words tend to be classified as perseverations (Galaverna et al., 2016; Henry and Phillips, 2006), regardless of the location of the words produced or their semantic integrity.

Other SVF scoring metrics focus on uncovering the process by which the sequences are generated. Semantic clustering is a technique of grouping words that are similar in meaning and belong to predefined subgroups (Ober et al., 1986; Raskin et al., 1992a; Robert et al., 1998). Troyer et al. (1997) proposed a formal analysis procedure that systematically identifies clusters of related words and assesses the number of switches between clusters as well as cluster size to allow for more fine grained analyses. For example, older people produce fewer words and switches than young people, but create larger clusters (Troyer et al., 1997; Troyer, 2000). While people with dementia normally produce fewer words than healthy controls, switching is impaired in frontal lobe related neurodegenerative diseases (e.g., Parkinson's disease) and cluster size decreases in temporal lobe involved neurodegenerative diseases (e.g., Alzheimer's disease) (Troyer et al., 1998b,a). Troyer's approach has been adapted widely to assess these scoring metrics, which are both thought to be necessary for successful SVF performance (Weiss et al., 2006; Lanting et al., 2009; Hurks et al., 2010; Zhao et al., 2013).

Given that SVF is one of the most widely used neuropsychological tests, normative data have been collected to establish expected SVF scores for healthy native speakers in several different languages. Examples include English (Tombaugh et al., 1999), Spanish (Benito-Cuadrado et al., 2002a), Arabic (Khalil, 2010a), Chinese (Feng et al., 2012), Dutch (Van Der Elst et al., 2006b), Hebrew (Kavé, 2005), Greek (Kosmidis et al., 2004a), and Persian (Ghasemian-shirvan et al., 2018). Normative studies will typically ensure that variation in performance according to age, gender and education is already accounted for in the published norms. Collecting normative data for different languages/ethnic groups is considered necessary given that SVF scores have been found to vary depending on the language/ethnic group the task is administered. For example, Kempler et al. (1998) found that Hispanic individuals produced significantly fewer animal names than Chinese, White, and Vietnamese groups and that Vietnamese individuals produced more animal names than Chinese, White, and Hispanic groups. These results have been linked to the length of the animal names spoken in each lan-

guage (e.g., the word "dog" in Spanish is perro, while in Vietnamese is chó); longer animal names result in fewer words generated. Therefore, it is important to establish normative data for the variations of SVF in different languages, determine what metrics have been used, and establish whether additional normative data or additional scoring methods for existing normative data would be beneficial. Both demographic information and cultural background can impact semantic memory organisation, category size and content (Rosselli et al., 2002; Strauss et al., 2006).Therefore, semantic categories included in SVF require scoring guidelines (Olabarrieta-Landa et al., 2017) and revealing culturally and linguistically unique retrieval strategies requires in-depth study of data from the population being considered (Olabarrieta-Landa et al., 2017). In this review, we focus on the Turkish language. Turkish is predominantly spoken by around 84 million people in Turkey (TUIK, 2022). There is also a large Turkish diaspora of more than 6.5 million all over the world, 5.5 million of whom live in Western Europe (Turkish Ministry of Foreign Affairs, 2022). Turkish differs substantially in word structure from more frequently studied languages, such as English. Turkish is a member of the Turkic language family, which also includes Azerbaijani and Turkmen. Its morphology is agglutinative (Durrant, 2013). Information such as case of nouns, plural/singular, or tense is indicated by adding an affix to a root word (Istek and Cicekli, 2007).

This systematic review addresses three research questions:

RQ1: What SVF tasks are commonly used for Turkish?

RQ2: What are the most commonly used scoring metrics? To what extent are more complex scoring approaches, such as clustering and switching (Troyer, 2000), used?

RQ3: What normative studies exist for Turkish, and how do the tasks and metrics map onto the existing literature on SVF in Turkish?

4.3 Materials and Methods

4.3.1 Search Strategy

This review was registered on PROSPERO with the protocol number CRD42020201585 and passed the suitability examination on 27/10/2020. In our study, we used the PICOS structure to frame the review questions Richardson et al. (1995). The PICOS structure

of the study is summarised in Table 4.1. We searched Web of Science, PsycINFO, EMBASE, MEDLINE, and two Turkish databases, TR-Dizin (National Database Index by ULAKBIM, the Turkish Academic Network and Information Centre) and Yok-Tez (Databases of National Thesis Center of the Council of Higher Education).

Term	Description
Population	Healthy native speakers of Turkish aged 16 and over.
Intervention	Standard semantic verbal fluency test. All semantic categories are included.
Comparison	In addition to normative data, we include papers that compare groups of healthy native speakers and papers that compare healthy native speakers to people with a health condition, such as neurodegenerative diseases or a mental disorder. Studies of bilingual speakers should have a monolingual control group.
Outcome	Scoring metrics used for semantic verbal fluency.
Study	Experimental studies that include more than one participant.

Table 4.1: PICOS Specification of the Systematic Review Inclusion Criteria

The general Boolean search formula was *((semantic fluency OR category fluency OR verbal fluency OR semantic verbal fluency OR COWAT OR Controlled Oral Word Association Test OR animal fluency OR animal naming) AND turk*)*, applied to the Topic, Title, Abstract, Author and Keywords fields. Local database searches were performed first using the English version of the formula, then using the Turkish translation, removing the *turk at the end. English keywords and their Turkish equivalents are given in Table 4.2. An initial search was performed on October 5, 2020 and updated on August 09, 2023.While the Controlled Oral Word Association Test (COWAT, (Rodríguez-aranda and Martinussen, 2006)) assesses letter fluency, we included it as a search term because some studies use COWAT to refer to semantic verbal fluency (Uslu, 2012; Ersan, 2014) or as an umbrella term for verbal fluency (Kiraç et al., 2014).

4.3.2 Inclusion and Exclusion Criteria

We included studies that reported primary SVF data from a group of at least two native Turkish speakers, and that were available in digital form. SVF data should be collected using the standard paradigm, i.e., naming as many words as possible from a given category in a short time, typically 60 seconds. At least one group of participants should consist of healthy adults.

English	Turkish
Semantic verbal fluency	Semantik sözel akıcılık
Semantic fluency	Semantik akıcılık
Verbal fluency	Sözel akıcılık
Category fluency	Kategori akıcılık
Animal fluency	Hayvan akıcılığı
Animal naming	Hayvan sayma
Controlled Oral Word Association Test (COWAT)	Kontrollü Kelime Çağrışım Testi

Table 4.2: Turkish Translations of English Query Terms for Search of Turkish Databases

The exclusion criteria fell into three broad categories: population; methodology; and accessibility.

The population-related exclusion criteria were "study of children"; "no healthy control group"; "no monolingual control group"; and "potential effect on SVF, no control". We defined "study of children" as a study that only reported data from participants under the age of 16. However, studies that specified a minimum age of 15 or 16 years for their participants, and that reported an average age of 18 or older, were included (Bora et al., 2019; Yilmaz, 2014). Studies that only focused on people with medical conditions ("no healthy control group") or only on bilingual people ("no monolingual control group") were also excluded. Finally, we excluded studies that reported individuals with potential undiagnosed conditions that might affect their performance on the SVF such as coronary artery disease ("potential effect on SVF, no control") (Ünlü et al., 2013).

The methodology-related exclusion criteria were: "only letter fluency"; "composite results"; and "no results reported". Studies that performed only letter fluency, phonemic fluency, or COWAT tests without any SVF component, were excluded as "only letter fluency". Studies reporting "composite results" did not provide metrics on SVF performance for a specific category. Instead, scores were merged across SVF categories (Aydınoğlu, 2015; Kaya and Alpözgen, 2022; İpekten, 2018; Erol et al., 2012), participants completed tests that involve alternating categories (Erdogan, 2016), scores from letter fluency and SVF were mixed (Kılavuz Ören, 2020; Gultekin et al., 2017; Beşer, 2019), or the type of verbal fluency test was not specified (Kandemir et al., 2009; Cevik et al., 2016; Gücüyener et al., 1998; Karahan et al., 2021). If studies reported no verbal fluency results at all, they were classified as "no results reported" (Er, 2014; Midi et al., 2011; Akgün, 2010; Boyle et al., 2021).

The accessibility-related criteria were "abstract only", which means that no full paper was available, and "no online version available", which applied to studies, in particular Turkish theses, that were only available in printed form.

4.3.3 Reference Management

Database search results were exported as BibTeX or RIS format except for Yok-Tez, which contains all Master, PhD and medical proficiency degree theses for diploma accreditation in Turkey. Although this system allows searching and downloading articles, it is not possible to save search results as bibliographic information. Therefore, the first author manually input references for the studies downloaded from Yok-Tez into Mendeley. Screening and data extraction were conducted using Covidence (2013). Duplicates were initially resolved automatically by Covidence. Remaining duplicates were resolved manually by prioritising the source in the order of MEDLINE, Embase, PsycInfo, Web of Science.

4.3.4 Screening and Extraction

All studies were screened by two members of the review team (RYK, KK, MW, SM) in the abstract and full text screening stages. Studies that were only available in Turkish were reviewed by RYK and KK, who are native speakers of Turkish, while studies available in English were reviewed by RYK, MW, and SM. Reviewers used the note facility to highlight potential issues and record their reasoning behind the decisions made. Disagreements were resolved through discussion between members of the review team.

Nine publications (Özcan et al., 2016; Mutlu et al., 2021; Özçelik-Eroğlu et al., 2014; Çabuk et al., 2020; Kiraç et al., 2014; Sezikli et al., 2018; Töret and Özdemir, 2021; Uzgan et al., 2021; Yavuz-Demiray, 2011) are theses that were later published as a peer-reviewed article; for those studies, we only extracted data from the published version, which is more likely to be accessible to the international community.

The PRISMA diagram (Figure 4.1) documents the screening process. The final number of studies analysed is 121.

The reviewers jointly designed an extraction template that was designed with the research questions in mind. Extraction was done by RYK, supported by KK. SM and MW were consulted in case of questions.

For each study, we extracted three types of *general information:* Language of the

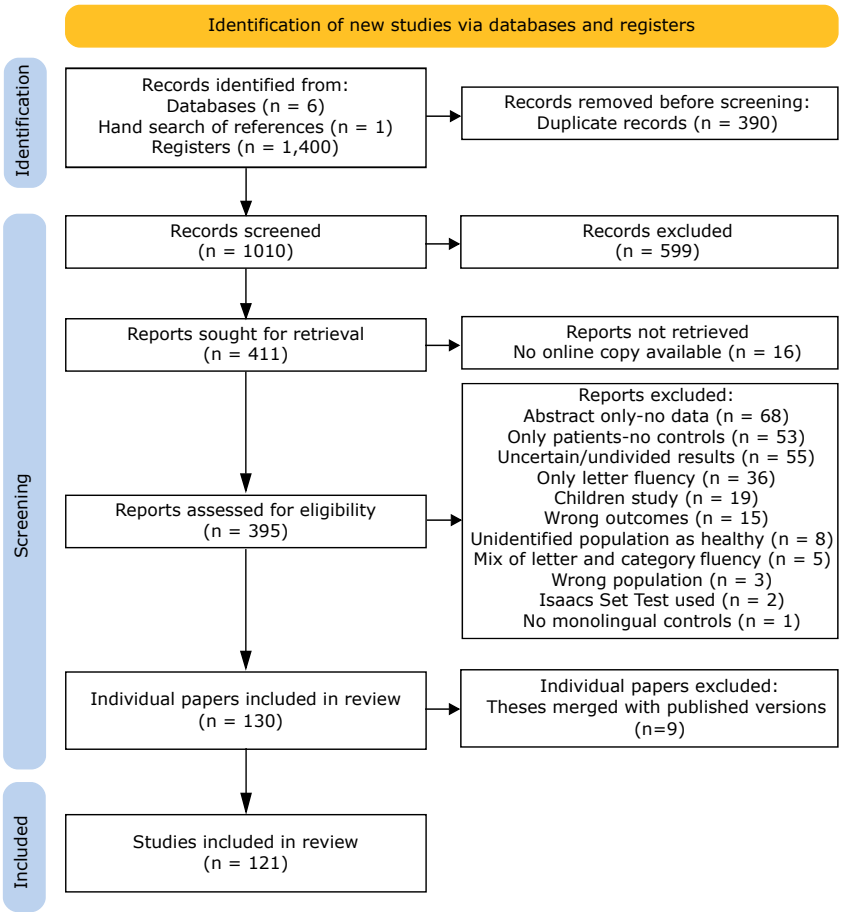


Figure 4.1: PRISMA flow chart

study (Turkish / English); Type of publication (thesis / journal article etc.); and Country in which the study was conducted (Turkey / Other). In order to establish commonly used tasks and scoring metrics (RQs 1 and 2), we extracted all semantic categories used (e.g., animals, vegetables), the word count metrics reported (mean / minimum / maximum / standard deviation etc.), any alternative analysis methods (clustering and switching / computational etc.), and the administration time of the SVF task. In order to distinguish between normative and non-normative studies (RQ3), we extracted *information about study type.* First, we determined the study design (longitudinal versus cross-sectional), and whether SVF was the focus of the study or part of a general assessment. If the study included a patient sample, we noted the type of health condition (e.g., cognitive impairment or dementia / mental illness / Parkinson's Disease etc.). If the study compared two or more groups of healthy participants, we noted how the groups were defined. Finally, we extracted demographic characteristics for the healthy native speakers of Turkish who were assessed. This included the age of the participants (older adults, age $\geq$64, adults aged 18-64, young people aged 16-25, and children / pediatric groups($\leq$16), and whether multiple age groups were differentiated. We noted the number of healthy participants, whether gender information was recorded, whether an indicator of level of education was used (years / levels / none), whether socioeconomic status was reported (No / Yes, occupation / Yes, income bracket), and age groups reported.

In the main results tables, we report the data extracted from each study and highlight missing information. The tables are summarised in the text using counts and percentages. For the tables, studies were split into normative and non-normative (c.f. Table 4.7). Non-normative studies were additionally split into studies comparing groups of healthy adults (c.f. Table 4.3), studies comparing healthy controls to people with a mental health condition (c.f. Table 4.4), studies comparing healthy controls to people with a neurodegenerative disease (c.f. Table 4.5), and studies comparing healthy adult controls to people with other diseases(c.f. Table 4.6).

4.4 Results

4.4.1 RQ1 and RQ2: Categories and Scoring Metrics

In general, animal naming is the preferred category in SVF, because it is least affected by country, cultural, educational and generational differences (Ardila et al., 2006). As

in other languages, for Turkish, animals are the most frequently used category (n=114, 94%). Other categories were first names (n=14, 12%), vegetables and fruits (n=6, 5%) and supermarket items (n=7, 6%). Even more rarely used categories included vehicles, clothes, parts of the body, furniture, famous people, breakfast items, food, beverages, and household items. Most of the rarer categories appeared in normative studies, which tended to provide verbal fluency data for a range of possible categories.

The administration time for SVF varied. Sixty seconds was used in 87% of studies (n=105), 12% of studies did not indicate the duration of the test (n=14), 2% used 90 seconds (Diker et al., 2016; Özdemir, 2015) (n=2), and one (1%) study with a dual task design used 30 seconds Özkul et al. (2021).

In terms of the metrics used, all studies reported the total number of words generated. Only four studies provided the number of words provided in each quarter of the 60 second period (Özdemir, 2015; Özdemir and Tunçer, 2021; Kandemir, 2006; Çukurova, 2020). In terms of more in-depth analyses, twelwe studies (Aki et al., 2022; Yazici, 2019; Çukurova, 2020; Uzgan et al., 2021; Çabuk et al., 2020; Kalafatoglu, 2015; Ersan, 2014; Sezikli et al., 2018; Sahin, 2022; Demiray and Ertan, 2023; Tumaç, 1997; İlkmen and Büyükişcan, 2022) (10%) reported perseverations, and six studies (Yazici, 2019; Çukurova, 2020; Çabuk et al., 2020; Kalafatoglu, 2015; Güneş et al., 2022; İlkmen and Büyükişcan, 2022) (5%) reported category violations. Clustering and switching was also used in only five studies (Çabuk et al., 2020; Kalafatoglu, 2015; Karaca, 2015; Uzgan et al., 2021; Altun, 2022) (4%). None of these papers provided a full translation for the original English taxonomy, although three papers listed the groups of animals used. Çabuk et al. (2020) and Kalafatoglu (2015) created a limited number of animal groups, namely farm, pets, forest, and zoological categories (insects, birds, and fish). Both studies also report perseveration. In their comparative study of frontotemporal lobe degeneration and semantic dementia, Karaca (2015) used different groups, namely poultry, ovine livestock, bovine livestock, forest, farm, birds, and insects. All studies were based on manual analysis, and differences between groups were established using standard inferential statistics. Bora et al. (2019) used manually established SVF scores as parameters for latent class analysis to explore group differences.

In the following subsections, we present an overview of the main tasks and metrics for non-normative studies. These fall into three categories: studies of bilingualism; comparisons between groups of healthy native speakers; and comparisons between patients and healthy control groups. Since a full meta analysis of Turkish SVF data

is beyond the scope of this review, and the number of studies across all categories is substantial, we only summarise findings for studies where SVF performance was a main focus.

4.4.1.1 Bilingualism

We found only one study that focused on speaking another language in addition to Turkish, which was a Turkish Masters dissertation (c.f. Table 4.3). Yazici (2019) compared the executive functions of bilingual Kurdish-Turkish speakers from Eastern and Southeastern Anatolia (n=80) to that of monolingual Turkish speakers from across the country (n=80). Animal naming was used, and word count, perseverations, and category violations were reported. Participants had 9 years or more of education and the age range was 20-54 years. Bilingual individuals' language proficiency levels were not formally determined, and there was no information about proficiency in other languages. While the study did not focus on SVF performance, the author argued that cultural aspects, such as languages spoken, need to be accounted for in normative studies of executive function in Turkish.

4.4.1.2 Comparison Between Groups of Healthy Native Speakers

Excluding normative studies, we found 21 studies conducted with healthy individuals, which are summarised in Table 4.3.Only two studies were conducted outside of Turkey. These were studies of minorities in Denmark (Nielsen and Waldemar, 2016; Nielsen et al., 2012) carried out with the help of an interpreter, and did not focus on bilingual or multilingual skills. Five studies focused on people aged 50 years and over. The topics of these studies fell into five distinct categories:

4.4.1.2.1 Physical activity: Exercise may have a positive effect on cognitive abilities due to the acceleration of blood circulation and increase of oxygen delivery to the brain (Mandolesi et al., 2018). Three studies examined the overall effect of exercise on cognition and used SVF as part of a standard assessment battery (Gökçe et al., 2021): tennis, (Gökçe, 2020): fencing or swimming, (Yeniçeri, 2019): balance training).

4.4.1.2.2 Cognition: In their study of the effect of various forms of training, including dual task training, on reducing fall risk, Balcı (2016) used SVF as part of their assessment battery. Despite the title of the thesis, Bozdemir (2008) focused on estab-

lishing Turkish normative data for the Pyramid and Palm Trees test, a semantic dementia test developed by Howard and Patterson (1992). SVF was administered as part of a larger battery of tests for comparison purposes. In the only study to specifically highlight SVF performance, Talas (2009) studied the relationship between magical ideation, handedness, and verbal fluency. They found a negative correlation between magical ideation scores and SVF scores, where the higher the magical thinking, the poorer the SVF performance.

4.4.1.2.3 Cognitive reserve: Lifetime experiences, such as educational and occupational attainment, literacy attainment and the involvement in cognitive and socially stimulating activities, are thought to increase the efficacy of cognitive processing in older age (for a review, see Arenaza-Urquijo et al. (2015)). Ekin and Çebi (2021) examined cognitive reserve and emotion regulation in older people, while Yildirim and Ogel-Balaban (2021) studied older people's use of social media (e.g., Facebook). Çağıl İnal (2019) compared professional musicians with non-musicians to investigate the effects of active interest in music. In all three studies, SVF was used as part of a standard assessment. Nielsen and Waldemar (2016) studied the relationship between SVF and literacy of healthy older Turkish immigrants in Denmark. Two categories were used, animals and supermarket items. Illiterate immigrants performed worse than literate ones on animal naming, but equally well on the supermarket category.

4.4.1.2.4 Familial risk: Studies in this group include healthy individuals who are first-degree relatives of people with a mental health condition and are therefore considered to have familial risk factors compared to the general population. The conditions included are schizophrenia (Kapu, 2019; Aydın, 2017; Berberoğlu, 2018; Noyan, 2011; Gürses, 2009) and substance-induced psychotic disorder (Çukurova, 2020). Bora et al. (2019) used Latent Class Analysis (LCA) to classify the neurocognitive performance of euthymic children of parents with bipolar disorder. They found three groups of performance: severe impairment; intermediate impairment; and good performance. SVF word count was significantly lower for offspring in the severe and intermediate impairment groups, compared to healthy controls and the "good performance" group.

4.4.1.2.5 Other: Studies in this category do not share a common framework in terms of their focus. Albayrak (2015) investigates procrastination in young adults, using SVF as part of their assessment battery. Evlice (2016) reports the effects of

age, gender, and education on healthy participants' performance on a standard battery of neuropsychological tests. SVF performance is negatively correlated with age (where the older the age, the poorer the SVF performance) and positively correlated with education (where the higher the level of education, the higher the SVF performance). In their comparison of Turkish immigrants and Danish citizens, Nielsen et al. (2012) examined the cross-cultural applicability of SVF, but without a monolingual Turkish control group. They used supermarket items as the test category. Nielsen et al. found that Turkish immigrants with higher acculturation produced more items, while older immigrants produced fewer. Güneş et al. (2022) investigated SVF performance of medical school students in terms of sleep duration: low, normal, excessive. They revealed that sleep duration has no significant difference on SVF. In another study, Sahin (2023), compared video gamers, athletes, musicians and participants who do not engage in any activity and found no significant difference between groups.

General Information			Demographics				Semantic Verbal Fluency		Analysis			
Study	**Language/Type/Country**	**Is SVF Focus?**	**Participant N (M/F)**	**Age Min-Max (SD)**	**Education years Min-Max (SD)**	**Economic status**	**Categories**	**Duration**	**Word count metrics**	**Other metrics**	**Group Comparison**	**Topics**
Yazici 2019	Turkish/Thesis/Turkey	No	160 (80/80)	20-54	Min:9	N/A	Animals	Unknown	Mean, SD	perseveration, category violations	Monolingual(Turkish) Bilingual(Kurdish and Turkish)	Bilingualism
Gökce 2021	English/Journal Article/Turkey	No	48 (25/23)	18-50 (±11.1)	Min:15	N/A	Supermarket items	60 sec.	Median, min, max	N/A	Tennis players, sedentary people	Physical activity
Gökce 2020	Turkish/Thesis/Turkey	No	54 (25/29)	18-25	Mean:13.5	N/A	Animals, Vegetables and fruits, Supermarket items	60 sec.	Mean, SD	N/A	Three exercise groups: fencers, swimmers, sedentary focusing gender difference	
Yeniceri 2019	Turkish/Thesis/Turkey	No	50 (24/26)	18-30	N/A	N/A	Animals	60 sec.	Mean, SD	N/A	Cognitive functions with and without balance exercises	
Ekin 2021	Turkish/Journal Article/Turkey	No	80 (40/40)	60-80 (±3,92)	Min:8	Occupation	Animals	60 sec.	Mean, SD, median, min,max	N/A	neuropsychological tests demographics and reading habit	Non-physical leisure activity
Yildirim 2021	English/Journal Article/Turkey	No	70 (36/34)	55–84 (±8.59)	Mean:9.17 (±4.24)	N/A	Animals	60 sec.	Mean, SD, min, max	N/A	Elderly facebook users non-users	
Inal 2019	Turkish/Thesis/Turkey	No	59 (33/26)	18–40	Mean:16.1	N/A	Animals, Vegetables and fruits, Supermarket items	60 sec.	Mean, SD	N/A	Musicians non-musicians	
Nielsen 2016	English/Journal Article/Denmark	Yes	41 (10/31)	≥50	Min:0 Mean:5	N/A	Animals, Supermarket items	60 sec.	Mean, SD	N/A	literate vs illiterate immigrants in Denmark	Cognition
Bozdemir 2008	Turkish/Thesis/Turkey	Yes	181 (87/94)	18-80	Min:5 Max:13+	Middle-income class	Animals	60 sec.	Mean, SD	N/A	Pyramids and Palm Trees Test semantic verbal fluency	
Talas 2009	Turkish/Thesis/Turkey	Yes	88 (44/44)	18-26 (± 1,52)	N/A	N/A	Animals, Vegetables and fruits, Supermarket items	60 sec.	Mean, SD	N/A	Right-handed, left-handed	
Balci 2016	Turkish/Thesis/Turkey	No	45 (6/39)	65-83 (±4,62)	Min:5 Max:13+	Occupation	Animals	60 sec.	Mean, SD	N/A	Dual task examination: cognitive tests with balance&walking training	
Kapu 2019	Turkish/Thesis/Turkey	No	60 (28/32)	18–58	Min:8 Max:18	Income bracket	Animals	60 sec.	Mean, SD	N/A	First-degree healthy relatives of schizophrenia patients vs healthy controls	Familial risk
Cukurova 2020	Turkish/Thesis/Turkey	No	82 (50/32)	18–64	Mean:9.85	Occupation	Animals	60 sec.	Mean, SD, quarters (e.g.15,30 sec.)	perseverations category violations	Siblings of substance-induced psychotic disorder	
Aydin 2017	English/Journal Article/Turkey	No	80	25-55	Min:5 Mean:9.01	Income levels	Animals	60 sec.	Mean, SD	N/A	Relatives of Schizophrenia patients	
Berberoglu 2018	Turkish/Thesis/Turkey	No	68 (30/38)	16-30	Min:11 Mean:14.39	Occupation	Animals	60 sec.	Mean, SD	N/A	Siblings of schizophrenic patients	
Noyan 2011	Turkish/Thesis/Turkey	No	62 (33/29)	15-35	Mean:13.94	Occupation	Animals	60 sec.	Mean, SD	N/A	First-degree relatives of patient with schizophrenia	
Gurses 2009	Turkish/Thesis/Turkey	No	92 (46/46)	≥18 (± 11)	Mean:12.3 (±3.11)	N/A	Animals, First names	60 sec.	Mean, SD	N/A	Relatives of schizophrenia	
Evlice 2016	Turkish/Journal Article/Turkey	No	100 (40/60)	18-77 (± 13.33)	5-15(±3.54)	N/A	Animals	60 sec.	Mean, SD, min, max	N/A	neuropsychological tests with demographic comparison	Other
Nielsen 2012	English/Journal Article/Denmark	No	109 (50/59)	50-87	Min:0 Max:16	N/A	Supermarket items	60 sec.	Mean, SD, min, max	N/A	Turkish immigrants Danish elderly	
Albayrak 2015	Turkish/Thesis/Turkey	No	124 (20/104)	20-49	Min:11 Max:17	Occupation and Income	Animals	60 sec.	Mean, SD, min, max	N/A	Cognitive functions with and without procrastination	
Güneş 2022	Turkish/Journal Article/Turkey	yes	50(23/27)	16-25	Min:13 Max:18	N/A	Supermarket items	60 sec.	Mean, SD	category violations	Low sleepers, Normal sleepers , Excessive sleepers	
Şahin 2023	Turkish/Thesis/Turkey	No	200(113/87)	18-22	Min:13 Max:19	N/A	Animals	60 sec.	Mean, SD	N/A	Video gamers, Athletes, Musicians, Do not engage in any activity	

Table 4.3: List of Studies that Compare Groups of Healthy Native Speakers with Full Extracted Data

4.4.1.3 Comparison Between Patients and Healthy Control Groups

Ninety-two out of 121 studies (76%) compared the performance of people with an impairment or condition to a healthy control group.

4.4.1.3.1 Mental Health Thirty-three of 92 studies (36%), summarised in Table 4.4, involved mental health conditions. In descending order of frequency, conditions included schizophrenia, bipolar disorder, obsessive compulsive disorder, depression, hyperactivity, borderline personality disorder, panic disorder, post-traumatic stress disorder, and schizotypy.

Only Sumiyoshi et al. (2014b) examined verbal fluency as a main focus of their study. Both verbal fluency types (semantic fluency and letter fluency) were investigated for schizophrenic patients in three languages: Japanese, Turkish and English. Like Turkish, Japanese is an agglutinative language, which uses affixes to indicate most relevant grammatical information. For Japanese and Turkish, performance was also compared to healthy controls. In Japanese, which uses two syllable-based scripts and one logographic script, letter fluency was more impaired than semantic fluency; in Turkish, which uses an alphabetic script, semantic fluency was more impaired than letter fluency. Uzgan et al. (2021) were the only authors to examine perseverations and number of switches in addition to SVF word count. There were no differences in any of the SVF measures between people with Obsessive-Compulsive Disorder and healthy controls. Unfortunately, the study, which used animals as a category, does not provide information on how the animal groups were formed.

4.4.1.3.2 Neurodegenerative Diseases Twenty-four of 92 studies (26%) studied neurodegenerative diseases (c.f. Table 4.5). Conditions that are frequently examined are the dementias, in particular Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), Parkinson's Disease, Multiple Sclerosis, and Essential Tremor.

Most studies used animals as their category, except for two studies which used the vegetables and fruits category. The data collection period was 60 seconds, except for Diker (2014) (90 seconds), Hanagasi et al. (2002); Yılmaz et al. (2020); Demiray and Ertan (2023) (no information), and Özkul et al. (2021) (30 seconds). Özkul et al. (2021) also used a novel scoring method, the number of correct words per second, which is an uncommon way of reporting SVF performance.

Five studies, which all use the standard animal category, focus on SVF. The longitudinal study by Dinç (2019), which compared MCI patients, MCI patients who

progressed to AD, and healthy controls, used the standard word count metric. They found that SVF was a good predictor of conversion to AD. The other three studies (Çabuk et al., 2020; Kalafatoglu, 2015; Karaca, 2015; Altun, 2022) used a clustering and switching approach, but defined their own categories. Çabuk et al. (2020) and Kalafatoglu (2015) created a limited number of animal groups, namely farm, pets, forest, and zoological categories (insects, birds, and fish). Both studies also report perseveration. In their comparative study of frontotemporal lobe degeneration and semantic dementia, Karaca (2015) used different groups, namely poultry, ovine livestock, bovine livestock, forest, farm, birds, and insects. Altun (2022) stated that the original Troyer animal groups were used, but did not provide information on how they were handled if a different animal name existed. The study was excluded perseverations for the total word count, but included for cluster analysis. If subcategories were said, the parent category name was excluded while creating clusters. For example, if ‘ ‘fish” was mentioned along with different fish names(salmon,sardine), the word ‘ ‘fish” was omitted.

Çabuk et al. (2020), comparing patients with amnestic MCI to healthy controls, found no difference in clustering and switching related parameters for SVF performance, but patients with amnestic MCI produced more perserverations. Kalafatoglu (2015) reported smaller SVF cluster sizes in patients with probable dementia compared to healthy controls. In Karaca (2015)’s study of patients with frontotemporal lobe degeneration, they produce fewer switches than healthy controls. Another study Altun (2022) focused on multiple sclerosis patients and found that multiple sclerosis patients produce significantly lower total words compared to healthy matches. No difference was observed between the groups in terms of mean cluster size, but fewer switches were created in the presence of Multiple sclerosis.

4.4.1.3.3 Other Conditions The remaining 35 (29%) studies, summarised in Table 4.6, covered six types of conditions, most of which are neurological. Ten studies focus on epilepsy, followed by addiction and stroke. All studies reported animal naming. Two studies added first names as a second category (Demir and Uluğ, 2002; Altunkaynak et al., 2019). In addition to word count, 3 studies report perseveration (Sezikli et al., 2018; Ersan, 2014; Sahin, 2022), while Kandemir (2006) reports performance every 15 seconds. SVF was always used as a supporting tool, never as a central focus of the research studies.

General Information			Participants						Semantic Verbal Fluency		Analysis	
Study	**Language/ Type/ Country**	**Is SVF Focus?**	**Clinical Case**	**Case N (M/F)**	**Healthy N (M/F)**	**Age min-max (SD)**	**Education Min-Max (SD)**	**Economic status**	**Categories**	**Duration**	**Word count metrics**	**Other metrics**
Balcilar 2019	Turkish/ Thesis/ Turkey	No	Bipolar disorder	37(16/21)	37(16/21)	18–64	Mean:12.20	N/A	Animals	60 sec.	Mean, SD	N/A
Bora 2019	English/ Journal Article/ Turkey	No	Bipolar disorder	71(30/41)	50(23/27)	15-30	Mean:13.37	N/A	Animals	Unknown	Mean, SD	Latent class analysis
Kiylioglu 2018	Turkish/ Thesis/ Turkey	No	Bipolar disorder	46(10/36)	43(5/38)	18–55	Min:5 Max:11+	Occupation, And Income	Animals	60 sec.	Mean, SD	N/A
Er 2015	Turkish/ Thesis/ Turkey	No	Bipolar disorder	82(31/51)	55(21/34)	18–65	Mean:11.9	N/A	Animals	60 sec.	Mean, SD	N/A
Esen 2010	Turkish/ Thesis/ Turkey	No	Bipolar disorder	55(21/34)	26(8/18)	18–64	Mean:12.19	N/A	Animals	60 sec.	Mean, SD	N/A
Bora 2007	English/ Journal Article/ Turkey	No	Euthymic bipolar disorder	65(30/35)	30(14/16)	18–55	Mean:12.02	N/A	Animals	60 sec.	Mean, SD	N/A
Uslu 2018	Turkish/ Thesis/ Turkey	No	Bipolar disorder, Major depressive disorder	24(12/12)	12(6/6)	≥18	Mean:11.6	N/A	Domestic animals	60 sec.	Mean, SD	N/A
Arat 2015	Turkish/ Thesis/ Turkey	No	Comorbid bipolar disorder, Attention Deficit Hyperactivity Disorder	101(42/59)	51(22/29)	18–64	Mean:12.9	Occupation	Animals	60 sec.	Mean, SD	N/A
Yilmaz 2014	Turkish/ Thesis/ Turkey	No	Attention Deficit Hyperactivity Disorder	30(16/14)	30(15/15)	16–65	Min:8 Max:22	N/A	Animals	60 sec.	Mean, SD, min, max	N/A
Senol 2017	Turkish/ Thesis/ Turkey	No	Borderline personality disorder	21(-/21)	20(-/20)	18–45	Min:8 Max:20	Occupation	Animals, First names	60 sec.	Median min, max	N/A
Dinc 2014	Turkish/ Thesis/ Turkey	No	Early onset depression, Late onset depression	30(11/19)	11(5/6)	≥60	Mean:9.5	N/A	Animals, First names	60 sec.	Mean, SD	N/A
Isikli 2011	Turkish/ Thesis/ Turkey	No	Major Depressive Disorder	18(6/12)	18(9/9)	29-58	Min:5 Max:18	N/A	Animals	60 sec.	Mean, SD	N/A
Kartal 2015	Turkish/ Thesis/ Turkey	No	Obsessive compulsive disorder	30(15/15)	30(15/15)	16–64	Min:5 Max:22	Occupation	Animals	60 sec.	Mean, SD	N/A
Begenen 2020	Turkish/ Thesis/ Turkey	No	Obsessive compulsive disorder	79(30/49)	41(19/22)	18–53	Min:5 Max:17	Occupation	Animals	60 sec.	Mean, SD, min, max	N/A
Tukel 2012	English/ Journal Article/ Turkey	No	Obsessive compulsive disorder	100(38/62)	110(43/67)	17–50	Mean:11.55	N/A	Animals	60 sec.	Mean, SD	N/A
OnderUzgan 2021	English/ Journal Article/ Turkey	No	Obsessive Compulsive Disorder	66(23/43)	75(29/46)	18–64	Mean:13.37	Occupation	Animals	60 sec.	Mean, SD	Perseveration, Clusters, switches (Troyer taxonomy)
Ozcan 2016	English/ Journal Article/ Turkey	No	Obsessive Compulsive Disorder	51(17/34)	21(12/9)	18–64(±10.8)	Mean:10.4(±4)	Occupation	Animals, First names	60 sec.	Mean, SD	N/A
Kurt 2013	Turkish/ Thesis/ Turkey	No	Obsessive compulsive disorder, Panic disorder	32(5/27)	26(5/21)	18-60	Min:5 Max:11+	Occupation	Animals	60 sec.	Mean, SD	N/A
Ekinci 2017	Turkish/ Thesis/ Turkey	No	Obsessive Compulsive Disorder, Panic Disorder, Generalized Anxiety Disorder	45(25/20)	29(9/20)	18–45	Min:5 Max:11+	Occupation	Animals	60 sec.	Mean, SD	N/A
Kilic 2013	Turkish/ Thesis/ Turkey	No	Panic Disorder	31(11/20)	31(11/20)	18–64(±8.6)	Mean:12.55 (±3.1)	N/A	Animals	60 sec.	Mean, SD, min,max	N/A
Eren-Kocak 2009	English/ Journal Article/ Turkey	No	Post-traumatic stress disorder	31	20	18–64	Mean:11.9	N/A	Animals, First names	60 sec.	Mean, SD	N/A
Ozak 2019	Turkish/ Thesis/ Turkey	No	Schizophrenia	130(72/58)	40(23/17)	18–55	Min:5 Max:11+	Occupation	Animals	60 sec.	Mean, SD, median, min, max	N/A
Aksoy-Poyraz 2011	English/ Journal Article/ Turkey	No	Schizophrenia	162(99/63)	51(32/19)	18–60	Mean:8.9	N/A	Animals	60 sec.	Mean, SD	N/A
Mutlu 2021	English/ Journal Article/ Turkey	No	Schizophrenia	46(28/18)	35(19/16	18–64	Mean:11.75	Occupation	Animals, First names	60 sec.	Mean, SD	N/A
Ozcelik-Eroglu 2014	English/ Journal Article/ Turkey	No	Schizophrenia	16(10/6)	8(5/3)	Mean:34.14 (±11)	Mean:11.16	N/A	Animals, First names	Unknown	Mean, SD	N/A
Ozcelik-Eroglu 2020	English/ Journal Article/ Turkey	No	Schizophrenia	16(10/6)	8(5/3)	Mean:34.14 (±11)	Mean:11.16	N/A	Animals, First names	Unknown	Mean, SD	N/A
Sevik 2011	English/ Journal Article/ Turkey	No	Schizophrenia	50(25/25)	25(11/14)	Mean:31.8	Mean:11	Occupation	Animals, First names	Unknown	Mean, SD	N/A
Sumiyoshi 2014	English/ Journal Article/ Turkey,Japan,USA	Yes	Schizophrenia	101(68/33)	50(34/16)	Mean:30.83	Mean:13.21	N/A	Animals	60 sec.	Mean, SD	N/A
Yazihan 2020	English/ Journal Article/ Turkey	No	Schizophrenia	25	27	Mean:22.68	Mean:11.9	N/A	Animals	60 sec.	Mean, SD	N/A
Zagli 2011	Turkish/ Thesis/ Turkey	No	Schizophrenia, Psychotic-like experience	51(20/31)	23(14/19)	15–65(Mean:41.88)	Mean:9.38	Occupation	Animals	60 sec.	Mean, SD	N/A
Canli 2011	Turkish/ Thesis/ Turkey	No	Schizotypy	22(12/10)	22(14/8)	15-18(±0.9)	Min:8	Occupation	Animals, First names	60 sec.	Mean, SD	N/A
Gürses 2022	Turkish/ Thesis/ Turkey	No	Schizophrenia, Psychotic disorder	40(23/17)	61(25/36)	18-60	Min:5 Max:19	Occupation, Income bracket	Animals	Unknown	Mean, SD	N/A
Özden 2022	Turkish/ Thesis/ Turkey	No	Major Depressive Disorder	30(10/20)	30(10/20)	Mean: 29.8 (±7.8)	Mean:15.45	Occupation	Animals, First names	60 sec.	Mean, SD	N/A

Table 4.4: List of Studies Comparing People with a Mental Health Disorder to a Healthy Control Group with Full Extracted Data

General Information			Participants						Semantic Verbal Fluency		Analysis	
Study	**Language/ publication type/ Country**	**Is SVF Focus?**	**Clinical case**	**Case N (M/F)**	**Healthy N (M/F)**	**Age min-max (SD)**	**Education Min-Max (SD)**	**Economic status**	**Categories**	**Duration**	**Word count metrics**	**Other metrics**
Hanagasi 2002	English/ Journal Article/ Turkey	No	Amyotrophic lateral sclerosis	20(15/5)	13(10/3)	Mean:9.2	N/A	N/A	Animals	Unknown	Mean, SD	N/A
Buyukgok 2017	Turkish/ Thesis/ Turkey	No	Apathetic and Non-apathetic Alzheimer's type Dementia	20(8/12)	9(5/4)	≥64 (Mean:71.26)	Mean: 8.84	N/A	Animals	60 sec.	Mean, SD	N/A
Cabuk 2020	English/ Journal Article/ Turkey	Yes	Mild cognitive impairment	20(6/14)	20(6/14)	48–83	Min:5 Max:11+	N/A	Animals	60 sec.	Mean, SD, min, max	Perseveration, Clusters and switches(Troyer taxonomy), Category violation
Dinc 2019	Turkish/ Thesis/ Turkey	Yes	Mild cognitive impairment, Alzheimer's disease	125(58/67)	134(53/81)	60–91 (Mean:72.6)	Mean: 10.83	N/A	Animals	60 sec.	Mean, SD	N/A
Kalafatoglu 2015	Turkish/ Thesis/ Turkey	Yes	Alzheimer's Disease	20	20	69-85(±3,31)	Min:0 Max:5	N/A	Animals	60 sec.	Mean, SD, min, max	Perseveration, Clusters and switches(Troyer taxonomy), Category violation
Karaca 2015	Turkish/ Thesis/ Turkey	Yes	Frontotemporal lobar degeneration, Semantic demans	21	21	48-71(±6,19)	N/A	N/A	Animals	60 sec.	Mean, SD, min, max	Clusters and switches(Troyer taxonomy)
Kayserili 2010	Turkish/ Thesis/ Turkey	No	Mild cognitive impairment, Alzheimer's Disease	61(24/37)	106(45/61)	≥50(±9.5)	Min:0 Max:11+	Occupation	Animals	60 sec.	Mean, SD	N/A
Kurt 2008	Turkish/ Thesis/ Turkey	No	Mild cognitive impairment, Subjective memory complaint	28(9/19)	16(5/11)	37-77 (Mean:58.65)	Mean:10.05	N/A	Animals	60 sec.	Mean, SD	N/A
Demirci 2018	Turkish/ Thesis/ Turkey	No	Duchenne Muscular Dystroph	58(-/58)	36(-/36)	26–67(Mean:40.8)	Mean:8.1	N/A	Animals	60 sec.	Mean, SD, median	N/A
BaysalKirac 2014	English/ Journal Article/ Turkey	No	Early multiple sclerosis, Clinically isolated syndromes, Definite diagnosis of multiple sclerosis	46(16/30)	40(14/26)	18–50 (Mean:32.55)	Mean:11.2	N/A	Animals	60 sec.	Mean, SD, min, max	N/A
Dogan 2020	Turkish/ Thesis/ Turkey	No	Essential tremor	40(20/20)	40(20/20)	18–60	Min:5 Max:11+	N/A	Animals	60 sec.	Mean, SD, Median, Min,Max	N/A
Sen 2014	Turkish/ Thesis/ Turkey	No	Essential tremor	11(7/4)	11(4/7)	18-55	Min:11 Max:14	N/A	Animals	60 sec.	Mean, SD	N/A
Ozkul 2020	English/ Journal Article/ Turkey	No	Multiple sclerosis	112(15/97)	25(5/20)	18–65	Mean:14.3	N/A	Vegetables and fruits	60 sec.	Median	N/A
Ozkul 2021	English/ Journal Article/ Turkey	No	Multiple sclerosis	60(16/44)	22(6/16)	18–64 (Mean:33.5)	Min:12 Max:16	N/A	Vegetables and fruits	30 sec.	Median	N/A
Tuncer 2012	English/ Journal Article/ Turkey	No	Multiple sclerosis	25(-/25)	17(-/17)	20–50 (Mean:34.33)	Mean:10.38	N/A	Animals	60 sec.	Mean, SD	N/A
Diker 2014	Turkish/ Thesis/ Turkey	No	Multiple sclerosis (Clinically isolated syndrome)	30(10/20)	20(7/13)	19–51 (Mean:31.82)	Min:5 Max:14 Mean:12.6	N/A	Animals	90 sec.	Mean, SD, min, max	N/A
Baran 2008	English/ Thesis/ Turkey	No	Non-Demented Parkinson's Disease	18(15/3)	18(12/6)	Mean: 64.52	Mean:10.30	N/A	Animals	60 sec.	Mean, SD	N/A
Degirmenci 2016	Turkish/ Thesis/ Turkey	No	Parkinson's disease	32(18/14)	32(12/20)	≥50 (Mean:59.9)	Min:5 Max:14	N/A	Animals	60 sec.	Mean, SD, min, max	N/A
Demirci 2010	Turkish/ Thesis/ Turkey	No	Parkinson's disease	27(17/10)	12(6/6)	Mean: 45.9	Mean: 9.8	N/A	Animals	60 sec.	Mean, SD	N/A
Yildirim 2012	Turkish/ Thesis/ Turkey	No	Parkinson's disease	39(31/8)	39	48-81	10,65 (±4.36)	N/A	Animals	60 sec.	Mean, SD	N/A
Yilmaz 2020	English/ Journal Article/ Turkey	No	Parkinson's disease	55(40/15)	20(9/11)	Mean: 65.27	Mean: 6.95	N/A	Animals	Unknown	Mean, SD	N/A
Öz 2022	Turkish/ Journal Article/ Turkey	Yes	Mild cognitive impairment, Alzheimer's Disease	144(62/82)	72(26/46)	Mean: 73.68	Mean: 10.49	N/A	Animals	60 sec.	Mean, SD	N/A
Altun 2022	Turkish/ Thesis/ Turkey	Yes	Multiple sclerosis	20(4/16)	20(4/16)	26-56 (Mean:39.9)	Min:5 Max:16	N/A	Animals	60 sec.	Mean, SD, Median, Min,Max	Clusters and switches(Troyer taxonomy)
Yavuz-Demiray 2023	Turkish/ Journal Article/ Turkey	No	Parkinson's Diseases, Essential tremor	60(37/23)	13(7/6)	40-80(Mean: 60.78)	Mean: 9.02	N/A	Animals	Unknown	Mean, SD, min, max	Perseveration

Table 4.5: List of Studies Comparing People with Neurodegenerative Disorders to Healthy Controls with Full Extracted Data

General Information			Participants						Semantic Verbal Fluency		Analysis		
Study	**Language/ publication type/ Country**	**Is SVF Focus?**	**Clinical case**	**Case N (M/F)**	**Healthy N (M/F)**	**Age min-max (SD)**	**Education Min-Max (SD)**	**Economic status**	**Categories**	**Duration**	**Word count metrics**	**Other metrics**	**Disease type**
Cakar 2020	Turkish/ Thesis/ Turkey	No	Epilepsy	64(21/43)	46(16/30)	18–64(Mean:31.28)	5-16 (Mean:12.68)	N/A	Animals	60 sec.	Mean, SD	N/A	
Evllice 2016	Turkish/ Journal Article/ Turkey	No	Epilepsy	30(15/15)	21(13/8)	18–64 (Mean:33.19)	Min:5	N/A	Animals	60 sec.	Mean, SD, min, max	N/A	
Kizil 2015	Turkish/ Thesis/ Turkey	No	Epilepsy	30(13/17)	30(13/17)	18–46(±5,93)	Min:5 Max:14	N/A	Animals	60 sec.	Mean, SD	N/A	
Sahin 2022	Turkish/ Thesis/ Turkey	No	Epilepsy	30(11/19)	15(5/10)	18-45 (Mean:27.38)	Mean:13.96	N/A	Animals	60 sec.	Mean, SD, median, min, max	Perseveration	
Celik 2015	Turkish/ Journal Article/ Turkey	No	Epilepsy, Psychogenic nonepileptic seizures	31(9/22)	20(5/15)	Mean: 29.5	Mean:10.6	Occupation	Animals	Unknown	Mean, SD	N/A	Epilepsy
Balcik 2019	Turkish/ Thesis/ Turkey	No	Juvenile myoclonic epilepsy	60(25/35)	30(13/17)	15-55(±8,5)	Min:5 Max:14	N/A	Animals	60 sec.	Mean, SD	N/A	
Cevik 2011	Turkish/ Thesis/ Turkey	No	Juvenile myoclonic epilepsy	20(6/14)	16(5/11)	16-40(±6,89)	5-15(±3,26)	N/A	Animals	60 sec.	Mean, SD	N/A	
Sezikli 2018	English/ Journal Article/ Turkey	No	Juvenile myoclonic epilepsy	45(11/34)	15(4/11)	Mean: 23.20	Min:5 Max:14	N/A	Animals	Unknown	Mean, SD	Perseveration	
Kilic 2015	Turkish/ Thesis/ Turkey	No	Mesial Temporal Lobe Epilepsy	28(16/12)	36(15/21)	18–55 (Mean:33.8)	Mean:9.5(±4)	Occupation and Income	Animals	60 sec.	Mean, SD	N/A	
Uslu 2015	Turkish/ Thesis/ Turkey	No	Mesial Temporal Lobe Epilepsy	79(30/49)	30(16/14)	18-57(±9,22)	Min:5 Max:14	N/A	Animals	60 sec.	Mean, SD, median,min,max	N/A	
Cengiz 2018	Turkish/ Thesis/ Turkey	No	Abnormal cranial imaging patients	30(14/16)	30(14/16)	18–50 (Mean:36.95)	5-16(±4,7)	N/A	Animals	60 sec.	Mean, SD, median	N/A	
Alibas 2017	English/ Journal Article/ Turkey	No	Acromegaly	42(19/23)	44(19/25)	18–64 (Mean:41.3)	Mean:8.08	N/A	Animals	60 sec.	Mean, SD	N/A	
Tutan 2019	Turkish/ Thesis/ Turkey	No	Chronic subjective tinnitus	60(38/22)	30(18/12)	21-60	5-16(±4,9)	N/A	Animals	60 sec.	Mean, SD, median,min,max	N/A	
Bakar 2010	English/ Journal Article/ Turkey	No	Hydrocephalus	15(9/6)	15(9/6)	18–64	N/A	N/A	Animals	60 sec.	Mean, SD	N/A	Neurological
Ulasoglu 2010	Turkish/ Thesis/ Turkey	No	Korsakoff amnestic syndrome	13(10/3)	13(10/3)	18–64 (Mean:39.3)	Mean:10.54	N/A	Animals	60 sec.	Mean, SD	N/A	
Buyukgok 2021	English/ Journal Article/ Turkey	No	Normal pressure hydrocephalus	30(13/17)	30(12/18)	≥60 Mean:67.72 (±8.13)	Mean:9.01 (±3.9)	N/A	Animals	60 sec.	Median, interquartile	N/A	
Eray 2013	Turkish/ Thesis/ Turkey	No	Restless legs syndrome	15(5/10)	15(5/10)	≥18 Mean:47.95	Mean:8.8(±4,1)	N/A	Animals	60 sec.	Mean, min, max	N/A	
Hatipoglu 2022	English/ Journal Article/ Turkey	No	Acromegaly	33(10/23)	30(8/22)	Mean: 46.6	Min:5 Max:16	N/A	Animals	60 sec.	Median, interquartile	N/A	
Ersan 2014	Turkish/ Thesis/ Turkey	No	Alcohol addiction	29(9/10)	30(9/11)	18–64 Mean:45.8	Mean:12.55	Occupation	Animals	60 sec.	Mean, SD	Perseveration	
Kocuk 2010	Turkish/ Thesis/ Turkey	No	Alcohol addiction	18(18/-)	18(18/-)	Mean:47.69	Mean:11.33	Occupation	Animals	60 sec.	Mean, SD	N/A	Addiction
Demir 2002	Turkish/ Journal Article/ Turkey	No	Alcohol addiction	34(34/-)	11(11/-)	Mean: 42.78	Mean:10.00	N/A	Animals, First names	Unknown	Mean, SD	N/A	
Çakmak 2022	Turkish/ Thesis/ Turkey	No	Opioid use disorder	93(93/-)	70(70/-)	19-48 Mean: 27.75	8-18 (Mean:11.0)	Income levels	Animals	60 sec.	Mean, min, max	N/A	
SenOzdemir 2020	Turkish/ Thesis/ Turkey	No	Neuromyelitis optica spectrum	20(3/17)	23(4/19)	23-54 Mean:38.05	Min:8 Max:16	N/A	Animals	60 sec.	Mean, SD, median,min,max	N/A	Autoimmune
BilginTopcuoglu 2018	English/ Journal Article/ Turkey	No	Sarcoidosis	21(7/14)	21(9/12)	18–64 Mean:43.80	Mean:7.59	N/A	Animals	60 sec.	Mean, SD	N/A	and
Altunkaynak 2019	English/ Journal Article/ Turkey	No	Behcet's Disease	30(20/10)	20(13/7)	Mean:33.07	Mean:11.74	N/A	Animals, First names	Unknown	Mean, SD	N/A	Inflammation
Kandemir 2006	Turkish/ Thesis/ Turkey	No	Infratentorial strokes	19(17/2)	19(5/14)	40-60 Mean:48.37	5-15 (Mean:7.3)	N/A	Animals	60 sec.	Mean, SD, Quarters (e.g.15,30 sec.)	N/A	
Erdal 2021	English/ Journal Article/ Turkey	No	Isolated cerebellar infarctions	23(17/6)	22(15/7)	18–64 Mean:53.56	Min:8	N/A	Animals	60 sec.	Mean, SD, median	N/A	Stroke
Temel 2020	Turkish/ Thesis/ Turkey	No	Thalamic hemorrhage	28(18/10)	28(18/10)	42-80 Mean:60.82	5-16(±3.80)	N/A	Animals	60 sec.	Mean, SD	N/A	
Sogutlu 2019	Turkish/ Journal Article/ Turkey	No	Memory complaints	56(15/41)	55(20/35)	≤55 Mean:36.45	Min:5 Max:14	N/A	Animals	Unknown	Mean, SD	N/A	
Akdemir 2021	Turkish/ Thesis/ Turkey	No	Chronic Pain	91(28/63)	70(22/48)	18–64 Mean:41.33	Mean:10.66	Occupation	Animals	60 sec.	Mean, SD	N/A	
Toret 2021	English and Turkish/ Journal Article/ Turkey	No	Blindness	20(14/6)	20(14/6)	≥18	Min:11 Max:16	N/A	Animals	60 sec.	Mean, SD	N/A	
Hangun 2022	Turkish/ Thesis/ Turkey	No	COVID-19	40(17/23)	40(17/23)	18-75 Mean:33.36	Min:5 Max:22	N/A	Animals	60 sec.	Mean, SD	N/A	Other
Iscen 2022	Turkish/ Thesis/ Turkey	No	Glucoserabrosidase mutation carriers	16(7/9)	18(8/10)	Mean:17.47	Min:30 Max:60	N/A	Animals	60 sec.	Mean, SD	N/A	
Cengiz-Al 2023	English/ Journal Article/ Turkey	No	Cavum septum pellucidum	26(9/17)	30(14/16)	18-50 Mean:36.95	5-16 Mean:10(±4,7)	N/A	Animals	60 sec.	Mean, SD	N/A	
Arisoy 2023	English/ Journal Article/ Turkey	No	Obstructive sleep apnea syndrome	78(70/8)	26(13/13)	18-50 Mean:38.57	Mean:11.97	N/A	Animals	61 sec.	Mean, SD	N/A	

Table 4.6: List of Other Studies Comparing a Group of Turkish Speakers with a Disease or Disorder to Healthy Controls with Full Extracted Data.

4.4.2 RQ3: Normative Studies

We found a total of seven normative studies that focus on SVF performance. Two further studies, Bozdemir (2008) and Evlice (2016), also collected data on SVF performance across a large sample of participants, but did not specifically aim to produce normative data for SVF. The details of the studies are given in Table 4.7 in order of publication year.

Except for Tuncer (2012), all studies recruited a balanced number of male and female participants. While Özdemir and Tunçer (2021) focused only on people aged 60 years and older, the other six studies recruited participants across the entire adult age range. All normative studies report level of education based on the Turkish national education system. Only Tuncer (2012) additionally explored illiteracy. No socioeconomic status information is reported, except (Tumaç, 1997). The number of participants varied between 58 (Özdemir and Tunçer, 2021) and 1431 (İlkmen and Büyükişcan, 2022).

In terms of category type, the most comprehensive study was conducted by Tuncer (2012), who collected SVF data for six different categories, namely animals, vegetables and fruits, vehicles, clothes, parts of the body, and furniture, from 400 participants. Tuncer (2012) was also the only study that focused on literacy. Two studies (Özdemir, 2015; Özdemir and Tunçer, 2021) examined five categories in addition to animals, namely breakfast items, famous people, food, beverages, and household items. Tuncer (2012) and Özdemir (2015) also list the most frequently used words and the first words produced for each category (e.g., dog, cat, horse, and donkey, for the animal category). While Şentürk (2019); İlkmen and Büyükişcan (2022); Tumaç (1997) focused only on animal naming, Aki et al. (2022) examined both animal names and first names.

Usually, the time allowed for participants to produce the words is 60 seconds. In the only study using 90 seconds (Özdemir, 2015), the total time was divided into 3 equal parts with 30 second chunks and the differences between the chunks were examined. This is equivalent to the common method, implemented by (İlkmen and Büyükişcan, 2022), of splitting 60 seconds into 15 second chunks and counting the number of words produced in each chunk. Word counts are reported excluding perseverations and category violations. Aki et al. (2022); Tumaç (1997) report perseverations and İlkmen and Büyükişcan (2022) report perseverations and category violations in addition to word counts. No other analysis approaches were found.

General Information			Demographics				Semantic Verbal Fluency		Analysis	
Study	**Language/Type/Country**	**Is SVF Focus?**	**Participant (N) Gender (M/F)**	**Age groups N(M/F)**	**Education groups N(M/F)**	**Economic status**	**Categories**	**Duration**	**Word count metrics**	**other metrics**
Tumaç 1997	Turkish/ Thesis/ Turkey	Yes	N:180 M:90 F:90	15-28 ages:60 32-45 ages:60 50-75 ages:60	0–5 years: 60 6-12 years: 60 12+ years: 60	Occupation	Animals	60 sec.	Mean, SD	perseveration
Tuncer 2012	Turkish/ Thesis/ Turkey	Yes	N: 382 M:170 F:212	18–24 ages: 78 25-34 ages: 89 35-44 ages: 36 45–54 ages: 51 55–64 ages: 64 65+ ages: 67	illiterate: 19 1–8 years: 87 9-11 years: 113 12+ years: 163	N/A	Animals, Vegetables and fruits, vehicles, clothes, parts of the body, furniture	60 sec.	Mean, SD	N/A
Ozdemir 2015	Turkish/ Thesis/ Turkey	Yes	N:120* M:60 F:60	18–29 ages: 30 30–44 ages: 30 45–59 ages: 30 60+ ages: 30	1-8 years:20/20 9-11 years:20/20 12+ years:20/20	N/A	Breakfast Items, Famous People, Food, Beverages, Household Items	90 sec.	Mean, SD, min,max, Each 30 sec.	N/A
Şentürk 2019	Turkish/ Thesis/ Turkey	Yes	N:200 M:99 F:101	18–29 ages: 71 30-39 ages: 60 40-49 ages: 69	5–8 years: 60 9-11 years: 66 12+ years: 74	N/A	Animals	60 sec.	Mean, SD, median, %5 percentile, %95 percentile.	N/A
Ozdemir 2021	Turkish/ Article/ Turkey	Yes	N:58 M:30 F:28	60-81 ages: 58	5–8 years: 20 9-11 years: 21 12+ years: 17	N/A	Breakfast Items, Famous People, Food, Beverages, Household Items	60 sec.	Mean, SD, Quarters(e.g.15,30 sec.)	N/A
Erden-Akı 2022	English/ Article/ Turkey	Yes	N:415 M:208 F:207	15-24 ages:46/42 25-34 ages:36/33 35-44 ages:30/36 45-54 ages:35/32 55-64 ages:31/34 65+ ages:30/30	5–8 years: 61/66 9-11 years: 66/62 12+ years: 81/79	N/A	Animals, Human names	60 sec.	Mean, SD	perseveration
Sohtorik-ilkmen 2022	English/ Article/ Turkey	Yes	N:1431 M:727 F:704	18-19 ages:60/60 20-29 ages:102/104 30-39 ages:103/103 40-49 ages:106/104 50-59 ages:105/102 60-69 ages:102/100 70-79 ages:92/84 80-89 ages:54/47	0–5 years: 159/160 6-12 years: 205/209 12+ years: 212/199	N/A	Animals	60 sec.	Mean, SD, Quarters Each 15 sec.	perseveration, category violations

Table 4.7: List of Normative Studies with Full Extracted Data. [*]Because our scope is adults, the results of the 30 children used in the study were excluded. Therefore, our assessment are based on 120 adults and the total number of people is 30 less than the actual number reported in the study.

Tumaç (1997), is an unpublished Masters dissertation, which reports normative results for a test battery that is sensitive to frontal lobe damage. Considering age and education levels together, 9 different groups were evaluated in the study based on a 3 age levels(young (15-18), middle-aged (32-45) and older (50-75)) and 3 education levels(low-medium-high) using animal naming category. In terms of total word count, the highest score was achieved by the middle-aged with high educated group Mean:26.65 SD:±5.04, while the old with low educated group Mean:19.25 SD:±4.46 and middle aged with low educated group Mean:19.4 SD:±3.66 resulted the lowest scores. The elderly low educated group created the highest number of perseverations, Mean:1.05 SD:±0.94, followed by the middle-aged high educated group, Mean:0.75 SD:±1.02. Although the occupational information of the participants was presented in a detailed list as background demographic information, it was not included in the analysis.

In Tuncer (2012)'s study, SVF scores were higher than letter fluency scores. The category yielding the highest scores was body parts, followed by fruits and vegetables, animals, clothes, vehicles, and furniture. Highly educated participants had higher scores; there was no difference with people who had only a few years of education and those who were illiterate. The 25-34 years age group had the highest scores, and the 65+ age group had the lowest.

Özdemir (2015) confirms Tuncer's results with regard to age and education in a sample of 150 participants aged between 15 and 81 years. In an in-depth study of the effect of education in older people aged 60-81, Özdemir and Tunçer (2021) showed that education level affected SVF scores, where the mean number of words every 15 seconds decreased linearly, but not chronological age.

Aki et al. (2022) report results from 415 participants. Age, gender, and level of education affected performance on SVF for first names. Higher scores on this task are linked to female gender, 9 or more years of education, and being in the 15-24 and 25-34 age groups. Animal SVF performance, in contrast, is only affected by level of education. Those with 12 or more years outperform those with 9-11 years of education, who in turn produce more words than those with only 5-8 years. There were not sufficient perserverations in either the animal or the names SVF tasks to warrant analysis. Şentürk (2019)'s study of the SVF performance of 200 healthy volunteers aged between 18 and 49 years confirms a strong effect of level of education with higher education associated with better SVF performance. In addition, male participants produced significantly more words than females.

İlkmen and Büyükişcan (2022) conducted the largest normative study in terms of number of participants with a total of 1431 healthy participants (M:727 F:704) from Istanbul sample. 8 different age groups (ranged 18 to 89) were compared in 4 education levels (elementary to postgraduate). They found that SVF performance has a significant correlation negatively with age but positively with education. Also they emphasised that there was no difference between genders. Considering narrow age groups has enabled the differences between groups to be shown as gradually increasing or decreasing, and it is seen that total word count decreases at the age of 50 and above. Additionally, the change in word production over time was gradually decreased regardless of age.

4.5 Discussion

To the best of our knowledge, this is the first systematic review focusing on available semantic fluency data for native speakers of Turkish. We reviewed normative and non-normative studies that report data on SVF frequencies that were produced in Turkish by healthy native speakers of the Turkish language, including studies where healthy native speakers formed a control group.

We found that, while SVF is a key part of neuropsychological assessment in studies of Turkish speakers, there is a paucity of normative data for all categories typically administered. In addition, one of the categories, first names, appears specific to Turkish. Unlike categories such as animals, fruits, or supermarket items, first names do not easily lend themselves to analyses that focus on the structure of a person's mental lexicon.

Most studies limit themselves to reporting word counts. We only found 15 (12%) studies that used more complex metrics. In cases where sequences were annotated with clusters and switches, the rules for scoring animal groups were not published, and the animal groups did not cover all of the groups defined by Troyer and collaborators (Kalafatoglu, 2015; Karaca, 2015; Çabuk et al., 2020; Uzgan et al., 2021; Altun, 2022). The lack of categorisation rules may explain why studies do not consider additional clustering and switching SVF metrics in Turkish studies. Moreover, without these rules, clinicians cannot consider clustering and switching in their patients' assessments. While word counts provide clinicians with a quick and easy measure, more detailed qualitative analysis for scoring SVF data can provide additional insights into human cognitive performance (Troyer et al., 1997; Mayr and Kliegl, 2000; Troyer,

2000; Abwender et al., 2001). Therefore, future work should provide normative data with the categorisation rules for the switching and clustering metrics in Turkish.

Our review revealed that there were very little data on the performance of bilingual versus monolingual speakers in Turkish, which is a clear gap in the literature given that 6.5 million Turkish speakers live in the diaspora abroad (Turkish Ministry of Foreign Affairs, 2022). Within Türkiye itself, several different languages are spoken, the most prominent of which is Kurdish. The Turkish Demographic and Health Survey from 2003 reports that 83% of Turkish people speak Turkish as their native language, 14% Kurdish, 2% Arabic, and 1% other languages (Koc et al., 2008). The only study in our data set that actually incorporates bilingualism involved Kurdish speakers in Turkey (Yazici, 2019). Two studies compared Turkish immigrants to Denmark to older native speakers of Danish (Nielsen and Waldemar, 2016; Nielsen et al., 2012), but they do not focus on bilingual competence.

Previous studies comparing bilinguals' and monolinguals' SVF performance in other languages have shown mixed results (Bialystok et al., 2008b; Luo et al., 2010; Paap et al., 2017; Sandoval et al., 2010; Patra et al., 2020). While monolinguals have been found to produce significantly more correct words than bilinguals (e.g., Sandoval et al. (2010)), the bilingual disadvantage is no longer found when the groups are matched for vocabulary (e.g., Bialystok et al. (2008b)). In a recent systematic review (Giovannoli et al., 2023), more than half of the studies did not report significant differences between monolinguals and bilinguals on SVF, a small number of studies demonstrated some bilingual difficulties but no study reported a bilingual advantage. As in our own review, most studies have assessed participants using the 'animal' category, with only three studies (Gollan et al., 2002; Keijzer and Schmid, 2016; Sandoval et al., 2010) adopting multiple categories. However, not all categories may be suitable for individuals living in different countries or speaking different languages; moreover, while animals may be the best cross linguistic category (Ardila, 2020), categories that draw upon everyday experience are better for assessing lower educated individuals (Nielsen and Waldemar, 2016).

When examining the provenance of the original 130 publications, we found that 82 studies (63%) were PhD or Master's theses, and only 48 studies (37%) were peer-reviewed papers. All theses except for Baran (2008) were written in English. Most of them were prepared by medical doctors during their specialization training in areas such as neurology or psychiatry. As noted in the Methods section, only 9 theses (Baysal Kiraç, 2012; Çabuk, 2018; Mutlu, 2018; Özcan, 2010; Sezikli, 2014; Töret,

2019; Özçelik-Eroğlu, 2012; Önder, 2019; Yavuz-Demiray, 2011) were published as peer-reviewed versions. Of the 48 peer-reviewed studies, 38 (79%) were published in English, nine (21%) in Turkish, and one (Töret and Özdemir, 2021) in both languages. Normative studies available in English were İlkmen and Büyükişcan (2022); Aki et al. (2022). This makes it very difficult for international researchers and clinicians to access data on SVF performance in Turkish.

4.5.1 Limitations

While we took great care to systematically search for studies beyond those produced by academic publishers (e.g., dissertations), other types of grey literature were not included. There was one normative studyfor which we were unable to obtain full text. Bingöl et al. (1994) is frequently cited as a key normative study, but no information about study design and results could be found. Yet, given our difficulty accessing Bingöl et al. (1994)'s study, it is unlikely that their normative data are available to other researchers and clinicians working in Turkish.

There were also a few terminological issues that made our search more difficult. For example, we found two studies that reported SVF as a subordinate task of COWAT, which is why it was included in the original search (Kiraç et al., 2014; Çakar, 2020). Ersan (2014) used "Word List Generation" for semantic fluency, and two studies (Ersan, 2014; Özcan et al., 2016) viewed category fluency as having two parts, animal naming and alternate fruit and animal naming.

Finally, due to the heterogeneity of the studies reported, we did not perform a meta-analysis of the SVF scores reported in the literature. Instead, we hope that researchers can use the groundwork laid in this paper to perform meta-analyses for smaller, well-defined groups of conditions.

4.5.2 Implications

The results of our systematic review indicate that individuals administering the SVF in non-clinical Turkish populations tend to use the animal naming category dominantly. Moreover, most studies simply report the total number of words produced. While a small number of studies provide more in-depth analyses that include clustering and switching, these studies do not provide the rules for scoring the various animal groups, making it difficult for individuals to apply these metrics to their own data. Therefore, there is a need for better Turkish normative data, which include the various semantic

categories administered, the various scoring metrics and the category grouping rules.

It should be noted that applying Troyer's scoring method manually is time-consuming as it requires careful reanalysis of the sequence of words produced; clinicians may have limited time to spend with their patients and so using additional scoring metrics is not feasible. Moreover, the established procedure specifically addresses animal naming, but offers insufficient subgroups for many animals (local animals, mythological animals etc.). Some names may be located in more than one group (e.g., parrot can be considered within the human use category as a pet or the zoological category as a bird), which makes cluster identification difficult (Woods et al., 2016a). Researchers have been investigating alternative approaches to identifying semantic clusters using generalisable, fast, and robust models that can be easily adapted to different languages and scenarios. Frequently investigated computational linguistics approaches are lexical relations (WordNet) (Quaranta et al., 2019; Paula et al., 2018), distributional semantics (word2vec, glove etc.) (Kim et al., 2019; Linz et al., 2017a; Holmlund et al., 2019; Voppel et al., 2021), and transformer language models (BERT) (Alaçam et al., 2022).

4.6 Conclusion

In summary, animal naming appears to be the preferred category when SVF is administered to Turkish-speaking individuals. Moreover, those administering the test appear to opt for the quick and easy method of scoring the SVF by simply considering the number of unique words generated. However, the SVF is considered a multifactorial test and so this score may not entirely represent an individual's performance; other scoring metrics such as cluster size and switching (Troyer et al., 1998b) may be differentially spared and impaired. Therefore, the methodology for producing semantic subcategories as well as normative data for the various SVF metrics is needed in Turkish to allow clinicians and researchers to understand their patients' cognitive profiles qualitatively.

The authors report that there are no conflicts of interest to declare

4.7 Role in Thesis

The findings from the systematic review were used to motivate the design of the data collection step (c.f. Chapter 6) and compare differences in SVF performance by gender and bilingualism status observed in our data set to findings from normative studies.

More importantly, our findings confirm that this thesis fills a clear gap in the neuropsychological literature on Turkish SVF. We found no reports of online collection of Turkish SVF sequences, no open source data sets of Turkish SVF, and a clear lack of computational linguistic approaches to the fine grained analysis of Turkish SVF sequences. Since the systematic review was limited to the psychological and psychiatric literature, we also checked for related papers in two important repositories of computational linguistics papers, arxiv.org and aclweb.org. Both sites were searched with the keywords "Turkish" and ("fluency" or "clinical"). The search yielded no additional papers on Turkish SVF from the field of computational linguistics. The only paper on Turkish clinical text processing we found was (Türkmen et al., 2023), which focused on the automatic analysis of radiology reports.

Chapter 5

Algorithm Validation: Semantic Verbal Fluency in Spanish-Speaking Colombian Alzheimer's Disease Patients

5.1 Overview

In this chapter, we aim to validate two computational linguistics approaches described in Chapter 3 on a Colombian Spanish dataset that consists of data from healthy controls and people with familial Alzheimer's disease. In particular, we investigate to what extent the two approaches can replicate manual clustering and switching analyses that follow the Troyer method (Section 2.4.2).

Our research questions are:

1. To what extent can the two computational approaches outlined in Chapter 3, the bigram and vector space method, detect the cluster structure of Colombian Spanish SVF sequences?
2. To what extent can cluster structure metrics derived from those two approaches replicate the differences observed in our dataset between healthy controls and Alzheimer's disease (AD) patients?

Many studies have found that people with neurocognitive diseases (e.g. Alzheimer's disease, Parkinson's disease) perform significantly worse on SVF tests than healthy participants in terms of total word count (McDowd et al., 2011; Tröster et al., 1998; Binetti et al., 1995; Auriacombe et al., 1993; Monsch et al., 1992). Such differences

have also been established using clustering and switching (Troyer et al., 1998b). Most studies show that AD patients make fewer switches than controls but that the groups do not differ in cluster size (Raoux et al., 2008; Bertola et al., 2014; Haugrud et al., 2011). Therefore, we expect that manually established cluster structure metrics will identify a difference between healthy controls and AD patients, which allows us to assess whether the automatic analysis methods preserve this distinction.

5.2 Methodology

5.2.1 Data

The clinical case group examined in this study comprised patients with familial Alzheimer's disease, a rare form of AD with a very early onset. The demographic characteristics and neuropsychological performance of members of the extended Colombian family affected by this condition have been extensively studied since they were first described by Lopera et al. (1997), including by the second supervisor of this PhD thesis MacPherson et al. (2012, 2015). The members of this family carry the single mutation E280A in the presenilin-1 gene (PSEN-1), leading to early-onset familial AD. Bekris et al. (2010) indicated that in 13% of early-onset AD cases, which are those in which symptoms begin to appear at approximately age 30, patients have inherited types of AD associated with three genes: APP, PSEN1, and PSEN2. Familial AD patients constitute approximately 1% of all AD patients according to report of Dementia UK (2020).

The dataset used in this study consisted of $N = 64$ 60-second SVF sequences collected from two groups of participants: healthy ($N = 50$) and AD ($N = 14$). The Animal category was used. At the time of compiling this chapter, detailed demographic information for the two groups of participants was unfortunately unavailable. This dataset was kindly provided by Sonia Moreno and Francisco Lopera at the Neuroscience Group, Sede de Investigaciones Universitaria (SIU), University of Antioquia, Medellin, Colombia.

5.2.2 Manual Annotation: Troyer Method

First, the original animal taxonomy given by Troyer et al. (1997) was translated into Spanish. Next, the animal groups were expanded with names produced by participants that were not featured in the original version of the taxonomy. The new taxonomy was created manually with the help of Charlotte Sudduth, a bilingual Spanish and English

speaker (see Acknowledgements 5.6). The translated and expanded Spanish version of the Troyer taxonomy can be found in Appendix A.7. In the second step, clusters and switches were generated for each person based on the adapted animal taxonomy, following the guidelines of the Troyer method.

5.2.3 Vector Space Model

The corpus used to create the vector space model (Word2vec) (see Section 3.5) was Spanish Wikipedia. Spanish has the ninth largest library in terms of number of articles. The Wikidump used contained 1,588,051 articles and was downloaded on 08/04/2020[1]. The computing infrastructure used was a standard Microsoft Windows laptop with a 2.9GhZ Intel Core i7-7500U CPU.

We followed the preprocessing sequence specified in Section 3.5. For stop word elimination, the list provided by Natural Language Toolkit (NLTK) (Bird and Loper, 2004) version 3.7 was used. We tested two lemmatisers for Spanish. **SpaCy** [2] is a well known library that supports 24 languages, including Spanish (Honnibal and Montani, 2017). **Pattern** [3] is a comparatively small but powerful (De Smedt and Daelemans, 2012) alternative. For stemming, we used the **Snowball** stemmer [4], an improved version of the Porter stemmer (Porter, 1980), as implemented in NLTK version 3.7.

Following Kim et al. (2019), we searched the hyperparameter space of Word2vec models for the best performing model. As discussed in Section 3.5, we created 12 models that varied according to three sets of hyperparameters:

- **Architecture:** CBOW and Skip-Gram,
- **Window size:** 4, 10
- **Dimensions:** 300, 600, 1000.

We used three thresholds for determining cluster boundaries based on the resulting cosine similarities: the first quartile, the mean, and the third quartile.

[1] Detailed information about the size and key features of the Spanish Wikipedia: `https://dumps.wikimedia.org/eswiki/`

[2] SpaCy lemmatiser tool page: `https://spacy.io/models/es`

[3] Pattern library page for Spanish: `https://www.digiasset.org/html/pattern-es.html`

[4] The library page of the Snowball Stemmer by NLTK: `https://snowballstem.org/algorithms/`

5.2.4 Statistical Analysis

5.2.4.1 Differences between Healthy Controls and AD Patients

The Mann-Whitney U test was used to assess the statistical significance of differences between the two groups. Tests were performed using the Python SciPy package via a statistical functions tools named scipy.stats [5], version 1.8.0.

5.2.4.2 Evaluation of Computational Models

We evaluated the performance of the bigram method and the trained vector space models using two metrics:

Correlation between predicted and actual parameters: Using Spearman's rho, (ρ), we assessed whether the number of switches and mean cluster size as determined by the computational methods were correlated with the corresponding parameters as established through manual annotation

Switch location: We assessed the prediction of switch location using precision (Eq.5.1), recall (Eq.5.2), and F1 score (Eq.5.3). In the equations, 'I' denotes the number of correctly determined switches (true positives), 'A' denotes the number of added switches (false positives), and 'D' is the number of deleted switches (false negatives). Figure 5.1 provides a worked example, and Figure 5.2 shows the definitions of I, A, and D in the form of a confusion matrix.

$$Precision = \frac{truepositives}{(truepositives + falsepositives)} = \frac{n(I)}{n(I)+n(A)} \tag{5.1}$$

$$Recall = \frac{truepositives}{(truepositives + falsenegatives)} = \frac{n(I)}{n(I)+n(D)} \tag{5.2}$$

$$F1score = 2 \times \frac{Precision \times Recall}{(Precision + Recall)} = \frac{n(I)}{(n(I)+n(A)) \times (n(I)+n(D))} \tag{5.3}$$

5.3 Results

5.3.1 Baseline: Analysis of the Spanish Dataset

Word Count and Perseverations: Healthy controls and participants with AD differ significantly in terms of total word count ($z_{(64)}$ = -5.254, $p < 0.000001$) and num-

[5]The library page of scipy.stats: : https://docs.scipy.org/doc/scipy/reference/stats.html

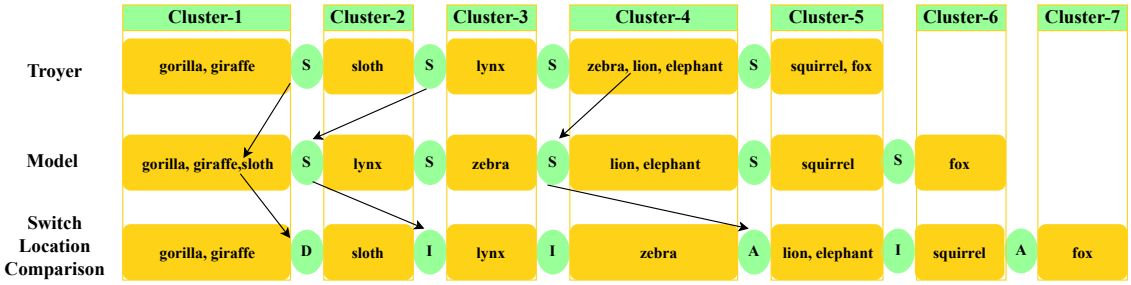


Figure 5.1: Comparing switch locations derived by Troyer method and vector space models. When the switch is in the same place for both computational and manual annotations, the switch comparison is marked with an 'I', because both locations are identical. When the switch occurs only in the computational annotation, it is marked with 'A', because the model added a new switch. When the switch is marked in the manual annotation but not in the computational version, it is marked with 'D', because the model deleted it.

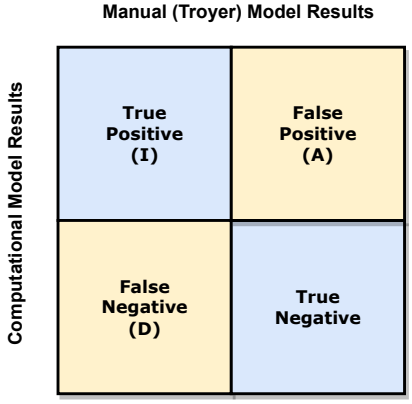


Figure 5.2: Confusion Matrix

ber of perseverations ($z_{(64)} = 4.149$, $p < 0.00001$). Full results are given in Table 5.1. Healthy control participants produced 21 words on average, as opposed to 11 for participants with AD. The AD group was also more likely to produce perseverations ($\overline{X}_{(14)} = 2.29$, $SD = 1.58$) than healthy participants ($\overline{X}_{(50)} = 0.46$, $SD = 0.73$). No participant produced any category violations.

Bigram Frequencies: In order to provide context for the performance of the bigram method, we report overall bigram statistics for the whole dataset. Overall, there are 1238 bigram tokens (frequency) and 966 bigram types (individual animal pairs) in the dataset. The most frequent animal pair was 'perro–gato' (dog–cat), which occurred 33 times, followed by 'león–tigre' (lion–tiger), which appeared 13 times. Of the total

								Mann-Whitney-U-Test			
Feature	**Participants**	**N**	**Mean**	**Mdn**	**Max**	**Min**	**SD**	**U-value**	**z-value**	**p-value**	η^2
Total words	AD	14	11.5	11	19	6	3.13	26	-5.254	1.39E-07***	0.657
	Healthy	50	21.16	21	30	11	4.11				
	All	64	19.05	20	30	6	5.59				
Perseverations	AD		2.29	2	5	0	1.58	605	4.149	5.88E-06***	0.519
	Healthy		0.46	0	3	0	0.73				
	All		0.86	0	5	0	1.24				

Table 5.1: Total words and perseverations. Mdn: Median, SD: Standard deviation, U: Mann Whitney U, z: z-statistic, η^2: estimate of effect size. ***: $p < .0001$)

number of animal pairs, n = 819 pairs (85%) were observed only once in the dataset, while n = 147 pairs (15%) were observed more than once.

								Mann-Whitney-U-Test			
Feature	**Participants**	**N**	**Mean**	**Mdn**	**Max**	**Min**	**Std**	**U-value**	**z-value**	**p-value**	η^2
Number of switches	AD	14	8.57	9	12	2	3.18	99.5	-4.06	4.50E-05**	0.507
	Healthy	50	13.24	13	22	7	3.2				
	All	64	12.22	12	22	2	3.73				
Mean Cluster size	AD		1.58	1.48	3	1.23	0.44	283	-1.08	0.279847	0.135
	Healthy		1.57	1.55	2.25	1.13	0.25				
	All		1.57	1.54	3	1.13	0.3				

Table 5.2: Troyer metrics are based on the number of switches and mean cluster size ($p-values$ are **$p < .001$).

Clustering and Switching Healthy people make significantly more switches than people with AD (healthy: $\overline{X}_{(50)} = 13.24$, $SD = 3.2$; AD: $\overline{X}_{(14)} = 8.57$, $SD = 3.18$, $p > 0.0005$), but we found no significant differences in the size of clusters produced (healthy: $\overline{X}_{(50)} = 1.57$, $SD = 0.25$; AD: $\overline{X}_{(14)} = 1.58$, $SD = 0.44$, $p < 0.3$). Detailed statistics are given in Table 5.2.

5.3.2 Evaluation of Computational Annotation

There was a strong correlation between the number of annotated switches from Troyer method and the number of predicted switches from the bigram method and all 36 vector space models (Table 5.3).

		Spearman Correlation Results									
		Hyperparameters			Snowball Stem		Pattern Lemma		Spacy Lemma		
Model name	**Th.**	*d*	**w**	**f**	***p*-value**	**rho(ρ)**	***p*-value**	**rho(ρ)**	***p*-value**	**rho(ρ)**	
Word2vec	**0.75**	1000	10	cbow	2.11E-14	0.7829	3.7E-15	0.7962	7.89E-16	0.8072	**
		1000	**10**	**skipgram**	4.79E-14	0.7764	6.71E-16	0.8083	2.15E-13	0.7637	**
		1000	4	cbow	1.14E-14	0.7878	8.78E-13	0.7511	2.13E-15	0.8002	
		1000	4	skipgram	3.9E-14	0.7780	9.02E-15	0.7896	2.76E-14	0.7808	
		600	10	cbow	8.78E-16	0.8065	5.84E-16	0.8093	8.14E-16	0.8070	*
		600	10	skipgram	1.69E-14	0.7847	4.51E-14	0.7769	8.97E-14	0.7712	
		600	4	cbow	2.74E-15	0.7984	2.29E-14	0.7823	1.3E-14	0.7867	**
		600	4	skipgram	1.61E-13	0.7662	7.15E-15	0.7913	7.11E-15	0.7914	
		300	10	cbow	3.35E-14	0.7793	4.81E-14	0.7763	3.8E-15	0.7960	
		300	10	skipgram	1.38E-13	0.7675	1.15E-13	0.7691	7.96E-14	0.7722	
		300	4	cbow	4.85E-15	0.7942	2.23E-15	0.7999	1.76E-15	0.8016	***
		300	4	skipgram	4.09E-13	0.7580	3.8E-15	0.7960	4.99E-14	0.7760	
	0.50	1000	10	cbow	6.86E-09	0.6485	4.34E-09	0.6549	1.11E-08	0.6416	
		1000	10	skipgram	4.82E-07	0.5810	4.56E-10	0.6843	1.88E-10	0.6950	
		1000	4	cbow	2.43E-07	0.5930	2.38E-07	0.5933	2.64E-07	0.5916	
		1000	4	skipgram	7.49E-09	0.6473	1.16E-09	0.6725	2.91E-11	0.7161	
		600	10	cbow	6.44E-09	0.6494	6.3E-09	0.6497	1.34E-07	0.6031	
		600	10	skipgram	4.11E-08	0.6220	2.07E-09	0.6649	9.62E-10	0.6749	
		600	4	cbow	1.28E-07	0.6038	4.97E-09	0.6530	1.98E-07	0.5965	
		600	4	skipgram	7.25E-07	0.5736	2.77E-08	0.6280	4.74E-10	0.6838	
		300	10	cbow	3.43E-09	0.6581	2.71E-09	0.6613	2.6E-07	0.5918	
		300	10	skipgram	3.33E-09	0.6585	1.74E-10	0.6959	2.43E-09	0.6628	
		300	4	cbow	7E-08	0.6136	1.65E-08	0.6358	4.77E-08	0.6197	
		300	4	skipgram	1.68E-09	0.6677	2.53E-09	0.6622	2.06E-08	0.6325	
	0.25	1000	10	cbow	3.16E-05	0.4953	0.000384	0.4305	0.000394	0.4297	
		1000	10	skipgram	1.67E-06	0.5580	4.32E-06	0.5390	0.000259	0.4417	
		1000	4	cbow	6.84E-05	0.4767	1.19E-05	0.5175	0.000616	0.4166	
		1000	4	skipgram	6.84E-06	0.5295	2.55E-05	0.5004	0.000072	0.4754	
		600	10	cbow	6.07E-07	0.5769	4.91E-06	0.5364	0.000335	0.4344	
		600	10	skipgram	1.38E-05	0.5143	6.55E-05	0.4777	0.000124	0.4614	
		600	4	cbow	5.54E-05	0.4819	2.09E-05	0.5049	0.0028	0.3676	
		600	4	skipgram	1.04E-05	0.5204	1.07E-06	0.5664	0.00034	0.4340	
		300	10	cbow	8.43E-06	0.5250	1.23E-05	0.5168	1.58E-05	0.5112	
		300	10	skipgram	5.37E-05	0.4827	1.28E-05	0.5159	0.000396	0.4296	
		300	4	cbow	3.33E-05	0.4941	0.000321	0.4356	0.000526	0.4213	
		300	4	skipgram	1.12E-06	0.5655	2.07E-06	0.5538	0.000161	0.4546	

Bigram	***p*-value**	**rho(ρ)**
	1.31E-11	0.725

Table 5.3: Correlation between number of switches for manual annotation and number of switches for computational models. All p-values are $p < 0.001$. Values for the best performing model are bolded. For the best performing threshold, the 3rd Quartile (0.75) scores are highlighted in green, with higher values represented by darker shades.

The bigram method achieved a correlation of ρ = 0.72, whereas the best vector space models achieved ρ ≥ 0.80. Overall, setting the threshold for a switch to the 3rd quartile of cosine similarities ($th = 0.75$) produced the best results. Visual inspection of Table 5.3 shows that using the Snowball stemmer consistently yields worse results than Pattern and Spacy.

The model 600_10_cbow performed consistently well, yielding the highest result with the Snowball stemmer (ρ = 0.8065) and the Pattern lemmatiser (ρ = 0.8093), and the second highest result with the Spacy lemmatiser (ρ = 0.8070). The best performing model with the Spacy lemmatiser was 1000_10_cbow, which achieved ρ = 0.8072.

For mean cluster size, on the other hand, the bigram method did not achieve a significant correlation between predicted and manually annotated cluster sizes, and neither do most of the 36 vector space models (Table 5.4). Only a few models achieved correlations that were significant at the $p < 0.05$ level when using the 0.75 threshold (3rd Quartile). The only models with a stronger correlation were 600_10_cbow with the Snowball stemmer (ρ = 0.332, $p < 0.01$) and 1000_10_skipgram with the Pattern lemmatiser (ρ = 0.344, $p < 0.01$). This suggests potential issues with switch placement.

We focused on the F1 score to evaluate switch location. The simple bigram method yielded $F1 = 0.756$. As before, the best vector space models outperformed it, with several models achieving $F1 > 0.8$ (Table 5.5). Spacy performed better than Snowball and Pattern, achieving the best overall F1 score. The best score for Pattern was $F1 = 0.8140$ (model 1000_4_skipgram), for Snowball $F1 = 0.8158$ (model 1000_10_skipgram), and for Spacy $F1 = 0.8309$ (model 1000_10_skipgram).

5.3.3 Replicating Baseline Results with Computational Methods

In this section, we examine to what extent annotations produced by computational models can replicate findings based on manual annotations.

For vector space models, we selected the model 1000_10_skipgram with a threshold of 0.75 and the Pattern lemmatiser, which tended to outperform the Snowball stemmer and the Spacy lemmatiser. While 600_10_cbow and 1000_10_cbow produced the best results in terms of the number of switches, 1000_10_skipgram and 600_10_cbow produced the highest scores for mean cluster size. In terms of switch location, models 1000_10_skipgram and 1000_4_skipgram performed best. When comparing the performance of all four models across tasks, we find that the model 1000_10_skipgram

					Spearman Correlation Results						
		Hyperparameters			Snowball Stem		Pattern Lemma		Spacy Lemma		
Model name	**Threshold**	***d***	**w**	**f**	**p value**	**rho**	**p value**	**rho**	**p value**	**rho**	
Word2vec	**0.75**	1000	10	cbow	0.063121	0.233671	0.057484	0.238711	0.091271	0.21286	
		1000	**10**	**skipgram**	0.022529	0.284845	0.005347	0.344246	0.057916	0.238311	*
		1000	4	cbow	0.282928	0.136279	0.3137	0.127939	0.060572	0.235903	
		1000	4	skipgram	0.033066	0.266819	0.020752	0.288568	0.032864	0.267115	
		600	10	cbow	0.00726	0.332531	0.016648	0.298329	0.12344	0.194551	*
		600	10	skipgram	0.025196	0.279701	0.011092	0.315537	0.088919	0.214386	
		600	4	cbow	0.078367	0.221653	0.23	0.152168	0.068394	0.229274	
		600	4	skipgram	0.028556	0.273838	0.007921	0.329109	0.018769	0.293055	*
		300	10	cbow	0.029055	0.273017	0.041585	0.255498	0.066833	0.230546	
		300	10	skipgram	0.02359	0.28274	0.049493	0.246594	0.050941	0.245092	
		300	4	cbow	0.053952	0.242078	0.04946	0.246628	0.0791	0.221124	
		300	4	skipgram	0.024432	0.281124	0.10592	0.20399	0.01844	0.293838	
	0.50	1000	10	cbow	0.428983	0.100598	0.338024	0.121712	0.509889	0.083884	
		1000	10	skipgram	0.899142	0.016161	0.439812	0.098266	0.532755	0.079418	
		1000	4	cbow	0.53628	0.078738	0.756795	-0.039474	0.651408	-0.05756	
		1000	4	skipgram	0.119397	0.196633	0.118927	0.196879	0.394408	0.108277	
		600	10	cbow	0.231989	0.151528	0.568927	0.07254	0.596959	0.06735	
		600	10	skipgram	0.562477	0.073751	0.648922	0.058003	0.413848	0.103914	
		600	4	cbow	0.449486	0.096211	0.99082	0.001467	0.878589	-0.01948	
		600	4	skipgram	0.52717	0.0805	0.19402	0.164481	0.315175	0.127553	
		300	10	cbow	0.160878	0.177368	0.20272	0.161365	0.881854	0.018949	
		300	10	skipgram	0.286568	0.135263	0.30028	0.131508	0.245594	0.147249	
		300	4	cbow	0.186146	0.167389	0.756072	0.039596	0.778983	0.035774	
		300	4	skipgram	0.567861	0.07274	0.252881	0.145024	0.712998	0.046878	
	0.25	1000	10	cbow	0.805732	0.031352	0.836363	-0.026333	0.451222	-0.09584	
		1000	10	skipgram	0.664968	0.055179	0.67619	-0.05322	0.251375	-0.14548	
		1000	4	cbow	0.926083	-0.01183	0.300623	0.131415	0.879776	-0.01929	
		1000	4	skipgram	0.704426	0.048344	0.956143	0.007012	0.244375	-0.14763	
		600	10	cbow	0.725862	0.044688	0.681712	0.052261	0.503582	-0.08513	
		600	10	skipgram	0.806561	0.031216	0.31949	-0.12643	0.407975	-0.10522	
		600	4	cbow	0.471842	0.091551	0.51898	0.082097	0.295521	-0.1328	
		600	4	skipgram	0.934108	0.010542	0.844389	0.025025	0.608333	-0.06528	
		300	10	cbow	0.943974	0.008961	0.389673	0.109359	0.957509	-0.00679	
		300	10	skipgram	0.856847	0.022999	0.86958	-0.020934	0.603988	-0.06607	
		300	4	cbow	0.60966	0.065034	0.38308	0.110878	0.689793	0.050862	
		300	4	skipgram	0.232048	0.151509	0.607379	0.065448	0.482689	-0.08933	

Bigram	**p value**	**rho**
	0.396303	0.108

Table 5.4: Correlation between mean cluster size for manual annotation and mean cluster size for computational models. The best performing model are bolded. The scores for the best performing threshold, 0.75 (3rd Quartile), are highlighted in green, with darker shades indicating higher scores.

					Switch Location									
		Hyperparameters			Snowball Stem			Pattern Lemma			Spacy Lemma			
Model name	**Threshold**	***d***	**w**	**f**	**precision**	**recall**	**F1 score**	**precision**	**recall**	**F1 score**	**precision**	**recall**	**F1 score**	
Word2vec	**0.75**	1000	10	cbow	0.753	0.82	0.7851	0.738	0.812	0.7732	0.8	0.841	0.8200	**
		1000	**10**	**skipgram**	0.793	0.84	0.8158	0.778	0.829	0.8027	0.81	0.853	0.8309	*
		1000	4	cbow	0.732	0.817	0.7722	0.741	0.802	0.7703	0.788	0.843	0.8146	
		1000	**4**	**skipgram**	0.786	0.829	0.8069	0.794	0.835	0.8140	0.8	0.835	0.8171	*
		600	10	cbow	0.766	0.83	0.7967	0.765	0.826	0.7943	0.775	0.832	0.8025	
		600	10	skipgram	0.774	0.839	0.8052	0.772	0.822	0.7962	0.796	0.835	0.8150	
		600	4	cbow	0.756	0.815	0.7844	0.751	0.801	0.7752	0.784	0.818	0.8006	
		600	4	skipgram	0.78	0.825	0.8019	0.79	0.821	0.8052	0.805	0.831	0.8178	
		300	10	cbow	0.783	0.822	0.8020	0.761	0.812	0.7857	0.775	0.824	0.7987	
		300	10	skipgram	0.76	0.82	0.7889	0.771	0.818	0.7938	0.785	0.832	0.8078	
		300	4	cbow	0.749	0.809	0.7778	0.761	0.811	0.7852	0.766	0.822	0.7930	
		300	4	skipgram	0.777	0.822	0.7989	0.765	0.815	0.7892	0.807	0.839	0.8227	**
	0.50	1000	10	cbow	0.835	0.568	0.6761	0.831	0.564	0.6719	0.853	0.581	0.6912	
		1000	10	skipgram	0.857	0.575	0.6882	0.863	0.578	0.6923	0.848	0.579	0.6881	
		1000	4	cbow	0.8	0.549	0.6511	0.794	0.546	0.6471	0.834	0.572	0.6786	
		1000	4	skipgram	0.881	0.595	0.7103	0.872	0.582	0.6981	0.865	0.582	0.6958	
		600	10	cbow	0.838	0.57	0.6785	0.791	0.556	0.6530	0.855	0.582	0.6926	
		600	10	skipgram	0.871	0.586	0.7006	0.863	0.573	0.6887	0.84	0.57	0.6791	
		600	4	cbow	0.801	0.551	0.6529	0.811	0.549	0.6548	0.83	0.561	0.6695	
		600	4	skipgram	0.854	0.574	0.6865	0.851	0.569	0.6820	0.865	0.583	0.6965	
		300	10	cbow	0.834	0.565	0.6736	0.804	0.565	0.6636	0.845	0.579	0.6872	
		300	10	skipgram	0.855	0.573	0.6862	0.847	0.565	0.6778	0.846	0.57	0.6811	
		300	4	cbow	0.802	0.55	0.6525	0.784	0.551	0.6472	0.814	0.554	0.6593	
		300	4	skipgram	0.853	0.573	0.6855	0.83	0.563	0.6709	0.836	0.565	0.6743	
	0.25	1000	10	cbow	0.856	0.289	0.4321	0.843	0.274	0.4136	0.874	0.293	0.4389	
		1000	10	skipgram	0.934	0.307	0.4621	0.909	0.293	0.4432	0.926	0.304	0.4577	
		1000	4	cbow	0.821	0.276	0.4131	0.828	0.271	0.4083	0.823	0.28	0.4178	
		1000	4	skipgram	0.933	0.304	0.4586	0.911	0.302	0.4536	0.915	0.303	0.4552	
		600	10	cbow	0.854	0.285	0.4274	0.856	0.281	0.4231	0.883	0.29	0.4366	
		600	10	skipgram	0.9	0.298	0.4477	0.909	0.293	0.4432	0.938	0.311	0.4671	
		600	4	cbow	0.831	0.277	0.4155	0.829	0.274	0.4119	0.824	0.276	0.4135	
		600	4	skipgram	0.923	0.306	0.4596	0.913	0.295	0.4459	0.923	0.308	0.4619	
		300	10	cbow	0.847	0.284	0.4254	0.833	0.275	0.4135	0.886	0.299	0.4471	
		300	10	skipgram	0.911	0.299	0.4502	0.896	0.285	0.4324	0.938	0.312	0.4682	
		300	4	cbow	0.822	0.272	0.4087	0.81	0.267	0.4016	0.82	0.28	0.4175	
		300	4	skipgram	0.908	0.304	0.4555	0.889	0.288	0.4351	0.93	0.307	0.4616	
Bigram					**prec.**			**recall**			**F1**			
					0.739			0.774			0.756			

Table 5.5: Comparison of switch locations between the Troyer method and different algorithms. The best performing methods are bolded.

performed consistently well across all switch location-related metrics. Also we chose Pattern lemmatiser as a morphological analyzer because compared to other two, Pattern reached better correlation scores for mean cluster size.

Both the bigram method (Table 5.7) and the vector space method (Table 5.8 successfully replicated the findings from the manual annotation, obtaining a significant difference in the number of switches produced and no differences in mean cluster size. Table 5.6 shows the descriptive statistics for the number of switches and mean cluster size, computed based on the output of the bigram method and the best performing

Model		hyperparameters			Number of switches				Mean Cluster size			
	Th	*d*	w	f	Mean	Mdn	Max	Min	Mean	Mdn	Max	Min
Word2vec	0.75	1000	10	cbow	13.44	14	22	3	1.43	1.4	2.2	1.07
		1000	**10**	**skipgram**	13.02	13	24	2	1.5	1.44	2.4	1.08
		1000	4	cbow	13.22	13	22	3	1.46	1.39	2.2	1
		1000	4	skipgram	12.84	12.5	24	2	1.51	1.44	2.25	1.07
		600	10	cbow	13.19	13.5	22	2	1.46	1.42	2	1.07
		600	10	skipgram	13.02	13	24	2	1.51	1.42	2.4	1.08
		600	4	cbow	13.03	13	22	3	1.48	1.42	2.2	1.07
		600	4	skipgram	12.7	13	23	2	1.54	1.46	2.29	1.07
		300	10	cbow	13.03	13	22	1	1.5	1.44	3	1.07
		300	10	skipgram	12.97	13	26	2	1.51	1.43	2.4	1.11
		300	4	cbow	13.02	13	22	2	1.49	1.41	2.2	1
		300	4	skipgram	13.02	13	24	2	1.5	1.42	2.4	1.07
	0.50	1000	10	cbow	8.3	8	17	1	2.34	2.17	4.75	1.23
		1000	10	skipgram	8.19	8	19	2	2.32	2.21	3.67	1.55
		1000	4	cbow	8.41	8	18	1	2.33	2.21	4.75	1.45
		1000	4	skipgram	8.16	8	16	2	2.35	2.22	4.6	1.56
		600	10	cbow	8.59	8	18	1	2.28	2.05	4	1.29
		600	10	skipgram	8.11	8	18	2	2.39	2.17	4.75	1.45
		600	4	cbow	8.27	8	17	1	2.4	2.19	4.75	1.38
		600	4	skipgram	8.17	8	20	1	2.36	2.28	4	1.33
		300	10	cbow	8.59	8.5	18	1	2.29	2.09	4	1.23
		300	10	skipgram	8.16	8	19	1	2.39	2.19	4	1.45
		300	4	cbow	8.59	9	17	1	2.29	2.09	4	1.33
		300	4	skipgram	8.28	8.5	18	1	2.38	2.18	6	1.33
	0.25	1000	10	cbow	3.97	4	9	0	4.99	4.04	12	2
		1000	10	skipgram	3.94	3.5	8	0	5.11	4	19	2.43
		1000	4	cbow	4	4	9	0	4.8	3.93	12	2
		1000	4	skipgram	4.05	4	9	0	4.64	4	12	2.12
		600	10	cbow	4.02	4	10	0	4.73	4.2	11.5	1.82
		600	10	skipgram	3.94	4	10	0	5.23	4	21	2.27
		600	4	cbow	4.03	4	11	0	4.72	4.1	12	2
		600	4	skipgram	3.95	3	10	0	4.97	4.22	12	2.22
		300	10	cbow	4.03	4	9	0	4.86	4.22	19	2
		300	10	skipgram	3.89	4	9	0	5.27	4.29	21	2.22
		300	4	cbow	4.03	4	10	0	4.8	3.98	12	2
		300	4	skipgram	3.95	4	11	0	4.85	4.22	16	2.3
Bigram					12.8	13	23	2	1.54	1.5	2.62	1.06
Troyer					12.22	12	22	2	1.57	1.54	3	1.13

Table 5.6: Descriptive statistics (Mean, Median (Mdn), Maximum (Max), Minimum (Min)) with the best morphological analyser, the Pattern Lemmatiser. The best models, which yielded the highest switch locations as seen in Table 5.5, are indicated with (*)$1^{st}\,place$,(**)$2^{nd}\,place$ and (***)$3^{rd}\,place$. Higher scores are represented by darker shades of green. Full descriptive statistics of other morphological analysers are given in Appendix A.8

vector space method.

								Mann-Whitney-U-Test			
Feature	**Participants**	**N**	**Mean**	**Mdn**	**Max**	**Min**	**Std**	**U-value**	**z-value**	**p-value**	η^2
Number of switches	AD	14	8	7	15	2	3.7	93.5	-4.157	3.1E-05**	0.52
	Healthy	50	14.14	13.5	23	7	3.86				
	All	64	12.8	13	23	2	4.63				
Mean Cluster size	AD		1.71	1.77	2.5	1.06	0.42	451.5	1.656	0.10076	0.207
	Healthy		1.49	1.46	2.62	1.14	0.26				
	All		1.54	1.5	2.62	1.06	0.32				

Table 5.7: Bigram method: differences between healthy controls and people with AD ($p-values$ are $^{***}p < .001$). Bigram results obtained using the selected best model, 1000_10_skipgram-0.75, with the Pattern lemmatiser.

								Mann-Whitney-U-Test			
Feature	**Participants**	**N**	**Mean**	**Mdn**	**Max**	**Min**	**Std**	**U-value**	**z-value**	**p-value**	η^2
Number of switches	AD	14	8.36	9	13	2	3.39	81	-4.36	1.22E-05**	0.545
	Healthy	50	14.32	14	24	6	3.84				
	All	64	12.86	13	23	2					
Mean Cluster size	AD		1.63	1.48	2.4	1.11	0.43	385	0.577	0.5748	0.072
	Healthy		1.47	1.42	2.12	1.08	0.22				
	All		1.51	1.44	2.3	1.1					

Table 5.8: Word2vec method: differences between healthy controls and people with AD ($p-values$ are $^{**}p < .001$). Word2vec results obtained using the selected best model, 1000_10_skipgram-0.75, with the Pattern lemmatiser.

5.4 Discussion

In line with the literature, we found significant differences between AD patients and healthy controls ($p < .001$), with greater total word count in healthy participants and higher frequency of perseverations in the AD group (see Table 5.1). The mean number of words produced per minute by AD patients in the E280A mutation group was found to be 10.56 by Lasprilla et al. (2003) and 10.91 by Rosselli et al. (2000b). In our study, the AD group had an average score of 11.5 words produced in a minute, while healthy

controls scored 21.16. Therefore, our study's overall word count aligns with the findings of these two previous studies. We also found that participants with AD produced more perseverations than matched healthy controls. Some studies have confirmed this difference in participants with sporadic (non-familial) AD (Butters et al., 1986; Tröster et al., 1998; Rosser and Hodges, 1994), although others have not (Butters et al., 1987; Tröster et al., 1989).

To the best of our knowledge, this is the first study using cluster-based metrics to examine the SVF performance of AD patients with the E280A mutation. Therefore, we compare our findings to those obtained in studies on people with sporadic (non-familial) AD. In line with Raoux et al. (2008), Bertola et al. (2014), Haugrud et al. (2011), and Troyer et al. (1998b), we found a lower number of switches in patients with AD. However, contrary to Troyer et al. (1998b), there was no statistically significant difference in mean cluster size between healthy participants and participants with familial AD ($Mean_{AD} = 1.58$ and $Mean_{Healthy} = 1.57$; see Table 5.2). Thus, even though the SVF dataset used was small and imbalanced (n = 64, of which 14 were AD patients), we were able to reproduce key findings from the literature.

Furthermore, we derived the same features through the algorithm-based methods. The metrics derived using the bigram method confirmed the manual analysis findings, but the vector space method also showed a small significant difference in mean cluster size between healthy controls and patients with AD ($p < 0.05$; see Table 5.8). This may be an artefact due to the number of significance tests conducted.

Both the simple bigram method and the more computation-intensive and complex vector space method were able to replicate the findings from manual annotation, indicating a difference in number of switches but not in mean cluster size. In order to select the best combination of hyperparameters and word representations for the vector space models, we assessed all possible combinations, following Kim et al. (2019). The best combination for Spanish was 1000_10_skipgram. The optimal combinations for Korean, as determined by Kim et al. (2019), were 1000_4_cbow and 600_10_cbow, and the best performing combination for English was 300_4_skipgram. English had a much larger Wikidump (20GB) compared to Korean (1GB). At 4GB, the size of the Spanish Wikidump is closer to Korean. We found that it was important to consider all three measures of model quality,namely switch location, number of switches, and mean cluster size. Several models that performed well for switch location and number of switches turned out to be unable to produce mean cluster sizes that correlated well with manual annotation.

This study has several limitations, most notably the small size of the underlying SVF dataset. Ideally, one would estimate the hyperparameters for the vector space model from a larger independent dataset of Spanish SVF sequences. It would also be interesting to see which of the parameters—vector length, window size, or embedding type—have the largest effect on performance. It would also be interesting to compare metrics such as mean cluster size and number of switches with the temporal variables suggested by Mayr (2002), since less switching can be a result of AD patients being stuck searching relevant cluster elements and therefore unable to continue with a new cluster. Understanding which clusters a person spends more time in, and which groups they have difficulty finding examples for, can provide a deeper understanding of the differences between patients and controls.

5.5 Conclusion and Further Research

In this study, we established that our implementations of the bigram method and the vector space method were capable of determining the fine internal structure of SVF sequences with sufficient accuracy, and that the resulting cluster structure metrics replicated significant differences between healthy controls and patients with AD. For this purpose, we analysed a unique dataset that included patients with familial AD. In Chapter 7, we adapted the successful methods for use with Turkish data, using the SVF dataset described in Chapter 6 for hyperparameter selection.

In terms of future work on familial AD, our analyses should be extended to include data from asymptomatic (healthy) genetic carriers (i.e. those who carry the PSEN-1, E280A mutation, but do not exhibit symptoms of the disease). Arango-Lasprilla et al. (2007) showed that asymptomatic carriers (Mean = 15.08) produced fewer words compared to healthy non-carriers (Mean = 17.45).

5.6 Acknowledgements

The Troyer animal taxonomy was adapted into the Spanish language with the help of Charlotte Sudduth.

Chapter 6

Online Collection of Turkish Semantic Verbal Fluency Data: Benefits and Challenges

6.1 Overview

In this chapter, we present an online spoken corpus of semantic verbal fluency data for the Turkish language with a corresponding lexical analysis. The traditional approach is to collect data during in-person meetings between the participant and practitioner via paper and pencil assessment (Connick et al., 2012; Dassanayake et al., 2021; Wall et al., 2017) or audio recording (Patra et al., 2020; Young et al., 2015; Pakhomov et al., 2015; Eekelaar et al., 2012). The limitation of these methods is that they require people to be in the same place at the same time. There are also studies that collect data through telephone interviews (Bunker et al., 2017; Diaz-Asper et al., 2021). These traditional data collection methods have two main drawbacks: (1) the need for more employees makes these methods labour-intensive, and (2) data collection through different channels may cause integrity and confidentiality problems regarding participants' personal information.

Moving away from these commonly used techniques, in this study, we developed a web-based data collection tool to reach participants. This approach allows participants to follow the instructions at their own pace, and also allows them to be a part of the study regardless of their surroundings. Participants only needed a device with internet access and audio recording capability, like a mobile phone or tablet. To the best of our knowledge, this is the first Turkish corpus collected with a web-based data collection

approach.

The aims of the study were to collect SVF data from healthy adults whose native language is Turkish and to present our findings regarding fundamental language characteristics in terms of different semantic categories. We will document our process in detail here and in the appendices to ensure that other researchers can replicate this study. We will also discuss the aspects of online data collection that can be replicated and the aspects that need improvement in order to increase uptake and data quality.

The abstract of the study was accepted for and presented at the 64th Annual Meeting of Psychonomic Society, which took place November 16–19, 2023 in San Francisco, California, USA. The abstract can be found in Appendix A.9.

6.2 Data Collection Design

We developed a two-stage web-based data collection environment which consisted of a participant questionnaire followed by a web-based audio recording application that recorded production of words in SVF categories. The survey directed the participants to the voice recording application through a URL. The audio recordings of the SVF responses and the participants' demographic information were linked via a unique participant ID assigned to each participant. The survey and application were first designed in English for the ethics and piloting steps, and were then translated into Turkish for distribution to the target audience.

The detailed framing of the data sample in this study was determined based on the systematic review of Turkish SVF studies in Chapter 4. Since the two-stage data collection process was carried out completely online, it was designed to be brief and engaging.

In the **questionnaire**, we focused on basic demographics: age, gender, and education level. These three demographic features are frequently included in Turkish normative SVF studies (Tuncer, 2012; Özdemir, 2015; Özdemir and Tunçer, 2021; Şentürk, 2019; Aki et al., 2022). We also asked about occupational status, a standard question that is not often reported in the Turkish literature, with the exception of some studies of healthy participants (Ekin and Çebi, 2021; Balcı, 2016; Çukurova, 2020; Berberoğlu, 2018; Noyan, 2011). Given the lack of data on bilingual speakers of Turkish demonstrated in the systematic review (c.f. Chapter 4), we included a dedicated section on bilingual abilities and ensured that the survey was also distributed in the Turkish diaspora.

In designing the **audio recording application**, we used the default test duration of 60 seconds, even though 30 sec.(Özkul et al., 2021) and 90 sec. (Özdemir, 2015; Diker, 2014) have also been reported. We selected three semantic categories: animals, vegetables and fruits, and supermarket items. These are the most studied categories in Turkish literature based on our systematic review in Chapter 4. We excluded first names since this category is rarely used in the literature.

6.3 Ethical Approval and Data Management

This study was approved by the School of Philosophy, Psychology and Language Sciences Ethics Committee at the University of Edinburgh on September 09, 2021 with the number 383-2021/2. See Appendix A.3.

The participant information sheet (PIS form) can be found in Appendix A.4. Consent was obtained using a yes/no question after the PIS text and included permission to share anonymised data with other researchers. The PIS provided detailed information about the nature of the study, the steps that needed to be completed, the rights of the participant, and data confidentiality. The PIS form was downloadable for participants' future reference. Each participant was provided with a unique participant ID. Contact information was provided in case participants wanted to exercise their right of withdrawal. No identifying information, such as name, date of birth, or IP address, was collected.

Survey data were stored on the Qualtrics survey platform hosted by the University of Edinburgh. The fully anonymous survey data was accessible only to the research team via log-in using EASE credentials. The data was transferred to password-protected, encrypted hard drives for analysis purposes and stored in a University of Edinburgh Microsoft SharePoint site.

SVF audio data were stored separately from the survey data. We collected three 70-second voice recordings from each participant during the SVF task. All audio files were in WAV format (waveform audio file). During data collection, audio files were stored by Google Cloud password-protected web hosting services accessible only to the research team. Afterwards, the audio data was transferred to password-protected, encrypted hard drives in order to be transcribed into written text. Transcribed data was no longer identifiable.

6.4 Data Collection Method

6.4.1 Participant Questionnaire

Figure 6.1 shows the survey process step-by-step. The survey began with demographic questions followed by language-related questions if the participant spoke at least one language other than Turkish. At the end of the survey, a unique 6-digit participant ID was displayed that was needed for the web app to link the SVF sequences and the survey information. The survey was implemented using Qualtrics XM Qualtrics (2005).

6.4.1.1 Demographics Block

Demographic information was collected with the sections described as follows:

Age group: To ensure anonymity, participants were asked to specify their age group: 18–29, 30–39, 40–49, 50–59, and 60+.

Gender: The possible choices for gender were 'Woman', 'Man', and 'Non-binary'. Participants had the right not to disclose their gender or to self-describe their gender.

Education level: Participants were asked about the highest level of education they had completed. If they were currently enrolled in university, it was recommended that they choose the highest degree received. The education levels provided as options were selected by considering the steps in the Turkish education system. Those are: elementary school (1st–5th grades), middle school (6th–8th grades), high school (9th–12th grades), associate degree (2 years of undergraduate), undergraduate (bachelor's) degree, postgraduate (master's) degree, and Postgraduate (PhD).[1]

Occupation status: The choices were 'Student', 'Homemaker', 'Employed full-time', 'Employed part-time', 'Self-employed', 'Unemployed', 'Retired', and 'Other, please describe'.

Diagnosis: We asked participants whether they had been formally diagnosed with any neurological or psychiatric conditions by a qualified medical professional. If the answer was yes, participants were asked to specify their diagnosis without using abbreviations.

Place of residence: 'Türkiye (previously Turkey)', 'United Kingdom', 'Germany',

[1]The division into grades follows the old education system, which all study participants living in Türkiye had followed. From the year 2012 onwards, the structure of compulsory education in Türkiye changed to four years of primary school, four years of middle school, and four years of high school (Gün and Baskan, 2014).

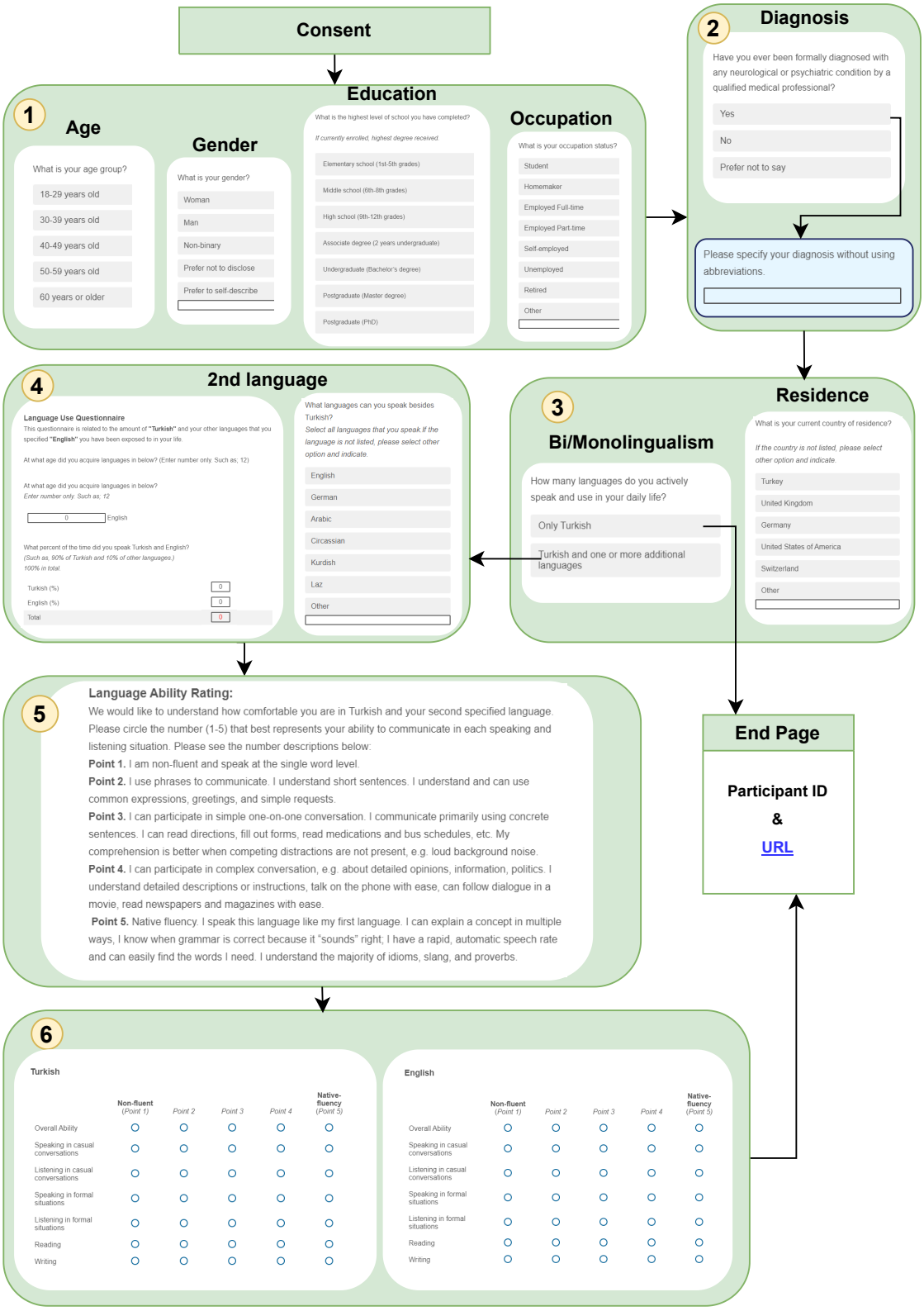


Figure 6.1: Steps of demographic survey.

'United States of America', and 'Switzerland' were provided as pre-defined choices, with the additional option to specify another country.

Language status: For multilingualism status, we asked how many languages participants actively used in their daily lives. If participants selected 'Turkish and one or more additional languages', the system directed the participant to the Bilingualism Block. Otherwise, the participant reaches the last page of the survey to be directed to the audio recording application.

6.4.1.2 Bilingualism Block

Bilingualism or multilingualism is usually measured with extensive validated tests. Comprehensive bilingualism assessments highlight many aspects of language acquisition and language use, such as exposure time to each language, language acquisition age, and preferred language use ratios in communication in different environments (family, friends, office, leisure activities). Since many native speakers of Turkish are bilingual which we recruited from the Turkish diaspora, we used a brief multilinguality assessment based on two validated questionnaires: the Language Experience and Proficiency Questionnaire (LEAP-Q) (Marian et al., 2007) and the Language and Social Background Questionnaire (LSBQ) (Anderson et al., 2018; Luk and Bialystok, 2013).

Spoken languages: We asked participants to list any other languages besides Turkish that they were proficient in.

Language use questionnaire (based on LEAP-Q): For each language, we asked participants about the age of language acquisition and the percentage of time the language was used. The percentages for all languages had to add up to 100%

Language ability scale (based on LSBQ): We asked participants to rate their proficiency in Turkish and the other specified languages in terms of four skills—speaking, listening, writing, and reading—and in formal as well as casual conversation situations. We used a five-point Likert scale, which was specified as follows:

- **Point 1:** I am non-fluent and speak at the single word level.
- **Point 2:** I use phrases to communicate. I understand short sentences. I understand and can use common expressions, greetings, and simple requests.
- **Point 3:** I can participate in simple one-on-one conversations. I communicate primarily using concrete sentences. I can read directions, fill out forms, read

medications and bus schedules, etc. My comprehension is better when competing distractions are not present (e.g. loud background noise).

- **Point 4:** I can participate in complex conversations about, for example, detailed opinions, information, and politics. I understand detailed descriptions and instructions, can talk on the phone with ease, can follow dialogue in a movie, and read newspapers and magazines with ease.

- **Point 5:** Native fluency. I speak this language like my first language. I can explain a concept in multiple ways. I know when grammar is correct because it 'sounds' right. I have a rapid, automatic speech rate and can easily find the words I need. I understand the majority of idioms, slang, and proverbs.

6.4.2 Audio Recording Application

The web-based audio recording application was co-designed by the three members of systematic review team (RYK, MW and SMcP) based on the results of the systematic review study presented in Chapter 4. The English version of an application was implemented by Danyi He as a part of her MSc degree requirement at the University of Edinburgh, under the supervision of MW and tutoring by RYK. Full text of the dissertation can be found on GitHub [2] to see the clear documentation of technical details, and the source code in GitHub repository [3] to download and employ the application. This thesis does not include the technical implementation of the application but provides key design elements, such as the SVF categories, storage, and time limitations.

6.5 Pilot Study for the Audio Recording App

The first step in developing the app was to build the minimum viable product, which is an early version of an application containing only sufficient fundamental attributes to allow it to be tested by users. We then conducted a pilot study in order to minimise and avoid potential drawbacks caused by the design. Piloting was carried out in English with five professional participants recruited from our personal research circle at the University of Edinburgh. Two of the attendees were informatics PhD candidates with knowledge of design and data analysis. The other three were PhD and postdoctoral

[2] Danyi He's MSc dissertation: `https://github.com/rykostas/DanyiHE_dissertation/blob/main/_Danyi_He_Dissertation.pdf`

[3] The code for the audio recording application: `https://github.com/rykostas/SVF`

researchers from the psychology department who were familiar with the SVF test and structure. Both groups were invited to use the application before a detailed explanation was given and were then asked questions about it. With the feedback received from the pilot study, improvements were made in the following aspects of the prototype design.

- **Design elements:** We made the navigation buttons larger and positioned them in the middle of the page, not aligned, so that they were much more visible. We highlighted certain words to draw attention to them. We made the system simple and uniform to ensure that the study can be used by people with basic computer skills. We also added sufficient but not excessive informative notes.
- **Instructions:** High-quality audio recordings facilitate the analysis step. Although background noise can be caused by many factors, like poor microphone quality or loud surroundings, and there is no certain way to get rid of it completely, we asked the users to take some precautions to deal with potential noise. These were:
  - Close the window and door
  - Move to a quiet room
  - Be sure there are no other sounds in the place or environment
  - Keep the microphone 20–30 cm away from your mouth, neither too close nor too far.
- **Preparation time:** A countdown to launch was added during the 10-second preparation time, with '3–2–1–start' written onscreen. With this addition, which keeps the user's attention active, our participants generally started when the recording began without pausing.

6.6 Distributed Version of the Audio Recording App

The translated Turkish version of the application was hosted on the Google Cloud Platform by the owner of the thesis, RYK. Participants logged into the system by entering the participant ID given to them by the survey system. The app only recorded participant's responses for the three assigned semantic categories. The recording order was animals, fruits and vegetables, then supermarket items. The audio recording application is illustrated in Figure 6.2 in detail.

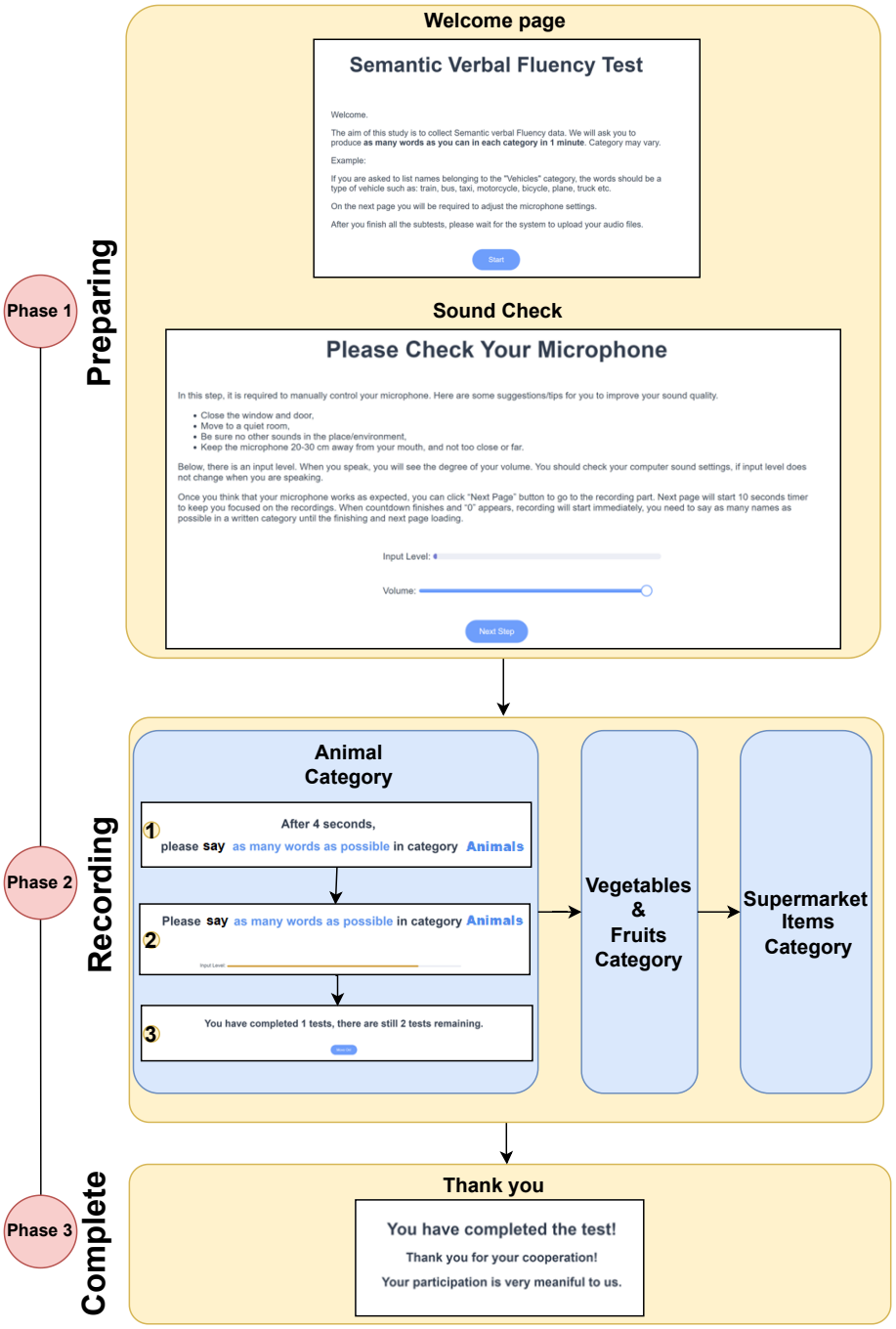


Figure 6.2: Steps of Audio recording application.

Participants followed the instructions and moved through the system at their own pace using the navigation buttons. Participant approval was needed at each step because there is no automatic jump between categories. When the participant was ready to begin the audio recording, the first category appeared and the participant was given 10 seconds of preparation time. At the end of that time, the person was given 70 seconds to produce as many words as possible in the relevant category. The remaining

time is hidden during the recording because in traditional data collection methods, the participant is not informed of the remaining time and is encouraged to continue producing words. The next category did not automatically start after the first category recording was completed. Rather, participant approval was requested again to give participants to rest. The name of the next category was not shared with the participant during this rest to ensure that the preparation time was a constant 10 seconds for everyone. In addition, advanced notice of the category may give participants time to research; as this test is about instant information recall, long preparation is not suitable for the nature of the test. The second and third categories were completed in the same way as the first.

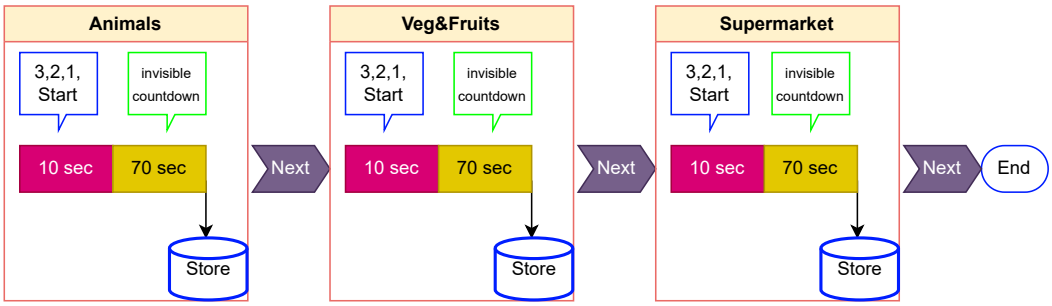


Figure 6.3: Flow chart illustrating the process of recording each category.

The SVF test time is usually 60 seconds, or sometimes 90, for each category. The reason for setting it to 70 seconds in this study was to counter potential effects of hesitation, breakdown, or panic. Due to the nature of the online data collection strategy, participants may also miss the first few seconds of the time, resulting in a test period shorter than desired. Therefore, 70 seconds was opted for to eliminate this problem and ensure a 60 second audio recording.

6.7 Recruitment

Participants were recruited through social media platforms such as Twitter (now X) and Instagram. The study was announced to the community of Turkish nationals living abroad with the help of the Turkish Consulate General Education Attaché Offices. Moreover, announcements were made on WhatsApp groups for the non-profit organisation ATAS (Association of Turkish Alumni and Students). In addition, we presented our study to many principals, teachers, and workers from government-funded state schools in Türkiye, and asked them to spread our research to their colleagues and so-

cial circles, either personally or via WhatsApp groups.

Participants volunteered for the study without receiving any compensation. Due to the privacy of the participants, no information was gathered about the region or province in which participants lived, only the country of residence was asked for.

6.8 Data Preparation

6.8.1 Automatic Transcription of Audio Files

Audio files in WAV format were transcribed using the Google Speech-to-Text API. The output was a text file for each participant. For an illustration of the process, see Figure 6.4. The API supports more than 125 languages; as our data is in Turkish we used 'tr' language code. We observed that participants began producing words promptly at the start of the recording, so we cut the last 10 seconds of all recordings. As a result, transcriptions were limited to the first 60 seconds. Transcriptions were checked manually at a later stage (c.f. Section 6.8.3).

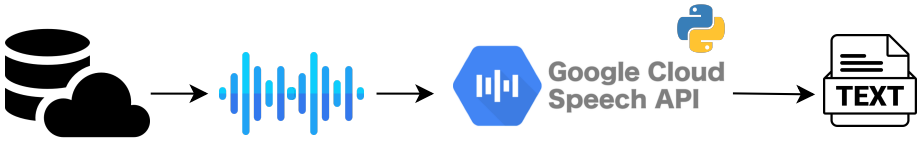


Figure 6.4: Process of Speech-to-Text

6.8.2 Pre-Processing

The initial raw texts were saved under the headings as 'Transcribed Text' and outputs were retrieved words from API. If the API was unable to produce any output, the file was coded as a 'Transcription Error' and resulted with **Empty** file. Empty files occurred for two reasons; either the participant chose to remain silent, or the sound could not be recorded due to microphone malfunction.

After automatic transcription of audio files, we manually controlled all transcriptions and conducted **Integrity Checks** to understand which files contained suitable data and increase the quality of transcriptions. The decision tree illustrates the whole integrity check process in Figure 6.5. While empty files are directly discarded, transcribed texts were examined under two subheadings: 'Adequate texts' and 'Errors'. Clear explanation of integrity check elements are as follows:

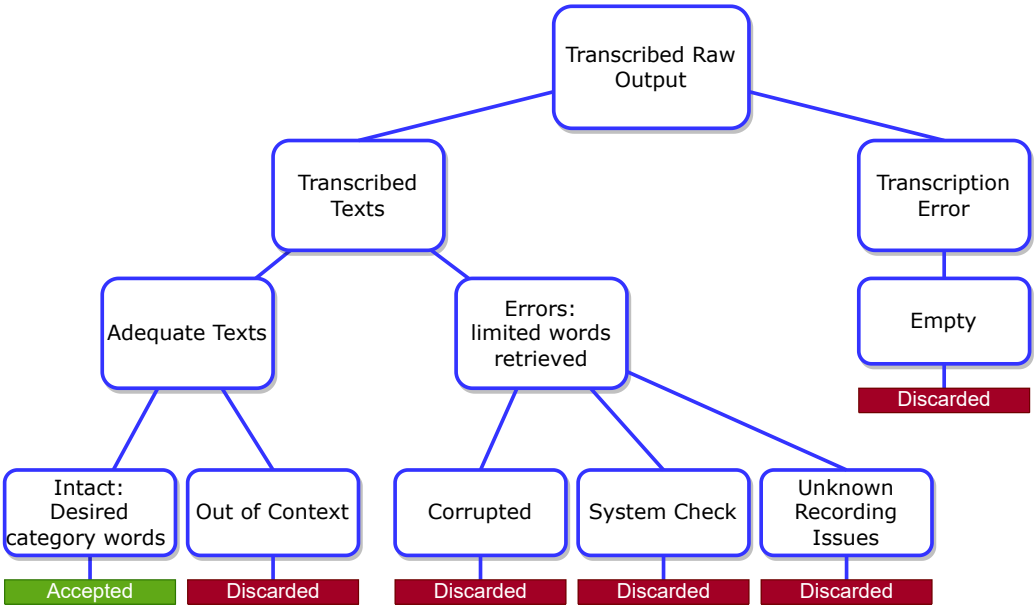


Figure 6.5: Integrity check process: Decision tree showing accepted and discarded transcribed audio recordings. The detailed numbers of integrity check elements in each category are given in Table 6.1.

Adequate texts: Enough words were retrieved, giving long outputs.

- ***Intact–*** Files containing words of the relevant/desired category (including perseverations) and sometimes thinking aloud or self-encouraging phrases. Also, recordings rarely included typos caused by Google API.
- ***Out of Context–*** Files consisting of sentences, more like storytelling than the sequence of words expected from participants. The participants' misunderstanding of the concept may underlie this error. Additionally, we observed few recordings in English rather than Turkish.

Errors: Limited words were retrieved, and the output is obviously missing length or contains unrelated words.

- ***Corrupted–*** Files that were unlikely to be fixed included those with unclear, crackling sounds or incomprehensible, interrupted words. In these cases, the API usually output these noises as random words. We encountered only a few files with this type of error which were irreversible or unfixable.
- ***System Check–*** During the data collection process, occasional system control entries were created by the admin to verify that the system was working properly.

- ***Unknown Recording Issues–*** Files containing errors that may have been caused by system design, the internet connection, or Google API services. The most obvious error encountered in this part was that half of the recording contained one category and the other half another category, such as 30 seconds of animals then 40 seconds of fruits and vegetables. Mixed category recordings may have been caused by a system crash or internet delay. There was also one file that was half empty.

After the integrity check, the numbers of acceptable recordings remaining for analysis were 105 for Animals, 101 for Fruits and Vegetables, and 101 for Supermarket Items. The detailed numbers for each category are given in Table 6.1. Accepted files were also examined in terms of **participant status**. There were two participant statuses. **(1) Alone:** Conducting the experiment unaided. $N_{Alone_{animal}} = 100$, $N_{Alone_{veg\&fruit}} = 101$, $N_{Alone_{supermarket}} = 97$. **(2) Assisted:** Completing the process with a helper. $N_{Assisted_{animal}} = 5$, $N_{Assisted_{veg\&fruit}} = 4$, $N_{Assisted_{supermarket}} = 4$. In this study, helpers were young adults who assisted participants who were older or had difficulties with the technology by giving them instructions. Since SVF is traditionally administered by a person, such records were not considered contrary to the nature of the test, but helpers had to avoid giving hints or suggestions. Acceptable verbal interventions were limited to encouraging directions such as 'keep going' or 'you are doing well'.

	All three categories	Animals	Fruit/Veg	Supermarket
Initial records	103(†)	137	123	116
Empty	5 (4%)	10 (7%)	5 (4%)	5 (4%)
Adequate				
• Intact	93 (68%)	105 (77%)	105 (85%)	101 (87%)
• Out of Context	6 (4%)	9 (7%)	7 (6%)	6 (5%)
- English	1 (1%)	1 (1%)	1 (1%)	1 (1%)
- Sentences	5 (4%)	8 (6%)	6 (5%)	5 (4%)
Errors				
• Corrupted	1 (1%)	3 (2%)	1 (1%)	1 (1%)
• System check	3 (2%)	3 (2%)	3 (2%)	3 (2%)
• Unknown issues	0	7 (5%)	2 (2%)	0

Table 6.1: Integrity check of transcribed audio files(N/%). 'All three categories' describes the recordings appearing in all three categories, since some people may exit the application without completing all categories. Therefore, 103† = 93 intact (accepted) + 5 empty + 6 out of context + 4 errors.

6.8.3 Annotation

In the pre-processing stage, we eliminated files unsuitable for further analysis. **Accepted** files were further examined manually in order to eliminate words outside of the required category and to create the final version of texts for analysis. All 'Transcribed Text' files were checked manually for potential mistranscriptions and misspellings. Some of the mistranscriptions resulted in phonetically similar word outputs that would have been counted as a category violation. Examples: 'koş' (run) instead of 'koç' (ram), 'yuva' (nest) instead of 'yılan' (snake), 'boşluk' (gap) instead of 'porsuk' (badger). Problematic transcripts typically occurred when SVF sequences were recorded with a helper or when there were clear voices in the background other than humming. However, most of the recordings were clear and transcribed correctly. Some typographical errors were kept which may be due to prosodic elements such as intonation, stress, or accent. For example, yılan (snake) in ID-743820, and pirana (piranha) in ID-110726 were misspelled in these instances, but spelled correctly for others.

Words were classified into no issue, perseverations, and category violations. In addition, we labelled helpers' speech and extra words produced by the participants while they were thinking aloud.

No issue: The SVF sequence included only words of desired category. If the category is animals, the participant listed animal names.

Perseverations: These are words that are repeated sequentially or intermittently. The number of perseverations per word is calculated using Equation6.1, and total perseverations were calculated with Equation 6.2.

$$Perseveration_{word_x} = N_{word_X} - 1 \tag{6.1}$$

$$Pers_{Total} = Pers_{word_x} + Pers_{word_y} + \ldots + Pers_{word_n} \tag{6.2}$$

Category violations: Any word that does not belong to the given category, like an animal name given when listing Fruits and Vegetables, is a violation. Three examples are given in Table 6.2. In the first example (ID-198916), the word is related to the animal world but is not an animal name. In other cases, the violations belong to other categories— Animals for the second example (ID-319968) and Supermarket Items (ID-493359) for the third.

Extra Speech—Thinking Aloud: There were a few instances where participants said additional phrases that appeared to be instances of thinking aloud while listing words from the desired category. Any extra speech between two words belonging to a

ID	category	SVF performance	Category violation (N)
198916	Animals	... cat, dog, bird, crocodile, hunting, food-chain, food, dolphin, whale, ...	3
319968	Veg&Fruits	... okra, peas, lettuce, radish, spinach, fish lettuce cucumber	1
493359	Veg&Fruits	... chickpeas, lentils, wheat, bread, sausage, salami, banana, kiwi, ...	3

Table 6.2: Category violation examples in the SVF outputs of three different participants are highlighted in orange. The sequences were translated into English. The number of category violations in the third column shows how many times violations occurred.

given category was counted as a single instance of extra speech, regardless of the word count. The three examples shown in Table 6.3 indicate that participants generally spoke to themselves when they could not remember any more words, when they remembered a word that they had said before, or when they tried to motivate themselves to continue. In some cases, it was limited to one (ID-758394) or two (ID-587309) instances of thinking aloud, but it happened up to seven times in rare recordings (ID-713964).

ID	category	SVF performance	Thinking Aloud (N)
758394	Animals	... cat, dog, There are no others this is all that comes to my mind, pigeon, magpie ...	1
587309	Veg&Fruits	... orange, nectarine, I said it mango, ..., onion, mushrooms, I don't think I can remember more, blackberry, mulberry	2
713964	Supermarket	..., lampchops, what else, ..., chocolate, I can't think of what else, snack, there is, chickpeas, ..., After that, drinks, there is, cola, ..., dishwasher tablet, I think I don't know, detergent, Yes, toilet paper	7

Table 6.3: Extra speech examples in the SVF outputs of three different participants are highlighted in orange. The sequences were translated in English. The number of extra speech instances indicates how many times these intrusions happened, not the number of words in the intruding sentences.

Extra Speech—Helper speaking: Any instance when the helper said something in order to encourage or praise the participant was counted as a single instance of extra speech. Table 6.4 contains three examples, one from each category. Generally, the aim of these intrusions was to support participants during recording and ensure successful task completion.

All words belonging to the desired category, including perseverations, were kept in the final SVF sequences. Extra speech, category violations, and helper speech were removed from the final SVF files. The detailed spreadsheet containing the number of each salient annotation element listed above calculated from the accepted records of each participant can be found in Appendix A.5.

Table 6.5 summarises how often each of the issues described in this section oc-

ID	category	SVF performance	Helper speaking (N)
233592	Animals	...horse, chicken, rooster continue, any other you remember, bird, seagull, fish tell me more, elephant, crocodile ...dinosaur awesome, try try, dog ...snake awesome, scorpion ...	5
954491	Veg&Fruits	...orange, lemon, lettuce, parsley any other, radish good, beetroot, carrot, persimmon, any other, apple, Cherry, sour cherry keep continue	4
240688	Supermarket	...tea, coffee, detergent any other, keep continue, think what else you buy, quickly, napkin, wipes	3

Table 6.4: Helper speech examples in the SVF outputs of three different participants are highlighted in orange. Sequences were translated in English. The number of instances of helper speech indicates how many times such intrusions happened, not the number of words in the intruding sentences.

cured in the dataset. While perseverations were very common, category violations were observed very rarely. Extra speech occurred in one out of five sequences. Five participants recorded sequences with the help of an assistant, which resulted in a total of 11 sequences with helper speech. For three of those five participants, the helper spoke in every SVF recording.

	All three categories	Animals	Vegetables&Fruits	Supermarket items
No issue	7 (8%)	35 (33%)	26 (25%)	46 (46%)
Perseverations	17 (18%)	52 (50%)	70 (67%)	46 (50%)
Category violations	0	2 (2%)	4 (4%)	0
Thinking Aloud (Extra speech)	6 (6%)	24 (23%)	20 (19%)	15 (15%)
Helper speaking (Extra speech)	3 (3%)	4 (4%)	3 (3%)	4 (4%)
Total Records	93	105	105	101

Table 6.5: Number of SVF sequences where the issue occurs, N(%). 'All three categories' describes the issues appearing in all three categories.

6.9 Overview of Dataset

6.9.1 Participants

The overall number of participants in both stages of the study, the demographic survey and SVF recording application, is shown in Figure 6.6. Participation in the survey was consented to by 286 people, but just 263 completed it by responding to all questions. After the demographic survey, 137 participants (52%) continued with the SVF data collection step, while approximately half of the participants (N = 126, 48%) did not proceed.

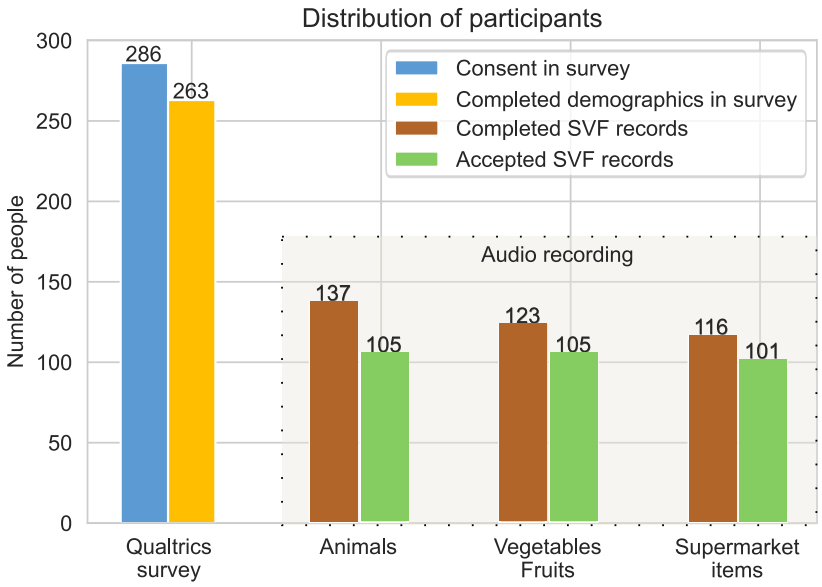


Figure 6.6: Participant numbers in the two stages of the study, the Qualtrics survey and SVF audio recording. Completed SVF recordings are the raw (initial) audio recording files that users saved. Accepted SVF recordings are the files used for analysis after exclusion of recordings that are incomplete, empty, or not meeting the requirements.

A total of 137 people started the SVF audio recording application; this is equal to the number of initial recordings for the Animals category. However, the number of recordings decreased to N = 123 (90%) for Fruits and Vegetables and N = 116 (85%) for Supermarket Items. Initial recordings were produced when participants started a task, but they did not necessarily result in a file, and the files may have been empty. Fourteen people dropped out between Animals and Fruits & Vegetables and seven left the app between Fruits & Vegetables and Supermarket Items. Possible reasons for dropping out may have included unwillingness to continue, system malfunction, or internet connection issues.

After the initial integrity check (c.f. Section 6.8.2), 32 Animals recordings (23% of 137), 18 Fruits and Vegetables recordings (15% of 123), and 15 Supermarket Items recordings (13% of 116) were eliminated. The number of accepted sequences for Animals was N = 105; it was N = 105 for Fruits & Veg and N = 101 for Supermarket Items. Some participants remained silent for one or both of the first tasks or produced SVF sequences that failed the integrity check. For this reason, we cannot assume that a person who produced a valid SVF sequence for Supermarket Items also produced valid sequences for Animals or Fruits & Veg categories.

						Accepted Records											
		Survey		App started		Animals		Veg&Fruits		Supermarket		One cat.		Two cat.		Three cat.	
	Group Features	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Age	18-29 age	62	23.57	28	20.44	18	17.14	19	18.1	16	15.84	2	16.67	3	30	15	16.13
	30-39 age	133	50.57	66	48.18	52	49.52	52	49.52	52	51.49	5	41.67	5	50	47	50.54
	40-49 age	40	15.21	23	16.79	19	18.1	18	17.14	17	16.83	4	33.33	1	10	16	17.2
	50-59 age	17	6.46	12	8.76	11	10.48	11	10.48	11	10.89	1	8.33	1	10	10	10.75
	60 age +	11	4.18	8	5.84	5	4.76	5	4.76	5	4.95	0	0	0	0	5	5.38
Gender	Male	121	46.01	66	48.18	51	48.57	51	48.57	50	49.5	7	58.33	5	50	45	48.39
	Female	138	52.47	70	51.09	53	50.48	53	50.48	50	49.5	5	41.67	5	50	47	50.54
	Non-binary	4	1.52	1	0.73	1	0.95	1	0.95	1	0.99	0	0	0	0	1	1.08
Education level	PhD	38	14.45	20	14.6	15	14.29	15	14.29	15	14.85	3	25	3	30	12	12.9
	Msc	84	31.94	38	27.74	29	27.62	29	27.62	30	29.7	3	25	2	20	27	29.03
	Bachelor degree	94	35.74	50	36.5	44	41.9	43	40.95	39	38.61	4	33.33	4	40	38	40.86
	Associate degree	10	3.8	4	2.92	3	2.86	3	2.86	3	2.97	0	0	0	0	3	3.23
	High(9-12 years)	13	4.94	8	5.84	7	6.67	7	6.67	6	5.94	2	16.67	0	0	6	6.45
	Secondary(6-8 years)	3	1.14	3	2.19	0	0	0	0	0	0	0	0	0	0	0	0
	Primary(1-5 years)	21	7.98	14	10.22	7	6.67	8	7.62	8	7.92	0	0	1	10	7	7.53
Occupation Status	Retired	7	2.66	6	4.38	4	3.81	4	3.81	4	3.96	0	0	0	0	4	4.3
	Homemaker	15	5.7	12	8.76	9	8.57	10	9.52	10	9.9	0	0	1	10	9	9.68
	Self-employment	4	1.52	3	2.19	3	2.86	2	1.9	2	1.98	1	33	0	0	2	2.15
	Full time employee	145	55.13	70	51.09	57	54.29	55	52.38	51	50.5	7	58	6	60	48	51.61
	Part time employee	4	1.52	3	2.19	3	2.86	3	2.86	3	2.97	0	0	0	0	3	3.23
	Student	56	21.29	32	23.36	20	19.05	23	21.9	22	21.78	2	16.67	3	30	19	20.43
	Unemployed	10	3.8	2	1.46	2	1.9	2	1.9	2	1.98	0	0	0	0	2	2.15
	Other	22	8.37	9	6.57	7	6.67	6	5.71	7	6.93	2	16.67	0	0	6	6.45
Place of residence	Germany	3	1.14	1	0.73	1	0.95	1	0.95	1	0.99	0	0	0	0	1	1.08
	USA	10	3.8	4	2.92	4	3.81	3	2.86	3	2.97	1	8.33	0	0	3	3.23
	United Kingdom	46	17.49	32	23.36	24	22.86	26	24.76	26	25.74	2	16.67	4	40	22	23.66
	Other	1	0.38	1	0.73	0	0	0	0	1	0.99	1	8.33	0	0	0	0
	Türkiye	203	77.19	99	72.26	76	72.38	75	71.43	70	69.31	8	66.67	6	60	67	72.04
Lang.	Monolingual	155	58.94	77	56.2	60	57.14	60	57.14	58	57.43	5	41.67	4	40	55	59.14
	Bi/Multilingual	108	41.06	60	43.8	45	42.86	45	42.86	43	42.57	7	58.33	6	60	38	40.86
Diagnosis	Yes	16	6.08	6	4.37	4	3.8	3	2.85	3	2.97	1	8.33	0	0	3	3.22
	No	238	90.49	127	92.7	98	93.33	98	93.33	95	94.05	10	83.33	10	100	87	93.54
	Prefer not to say	9	3.42	4	2.91	3	2.85	4	3.8	3	2.97	1	8.33	0	0	3	3.22
	Total	263	100	137	100	105	100	105	100	101	100	12	100	10	100	93	100

Table 6.6: Demographic distributions of participants at different stages of the data collection process. *Survey* includes people who completed the survey. *App started* includes people who started the SVF recording application. The *Animals*, *Fruits&Veg*, and *Supermarket Items* columns provides details of the people who provided valid recordings for the corresponding category. The *One*, *Two*, and *Three category* columns include those who provided valid recordings for the corresponding number of categories. Valid records refer to accepted SVF sequences that have passed the integrity check process.

6.9.2 Demographic Characteristics

Detailed participant characteristics are given in Table 6.6. In the table, we chart participant drop off between survey completion and the start of the SVF app. The table

gives the demographics of those who produced an accepted recording for each of the three tasks. Finally, we report on those who completed only one SVF task, only two SVF tasks, or all three SVF tasks. Note that participants who completed one of the later tasks successfully may have produced incomplete files earlier.

While N = 263 people completed the survey questions, only N = 137 people continued to the SVF application and completed at least one of the categories.

Most participants were in the 30–39 age group, which was followed in prevalence by the 18–29 age group. Participants were generally highly educated and tended to be students or employed full time. Many groups were substantially underrepresented. For example, only three people identified secondary education as their highest level of education, and none of them completed the SVF tasks, even though they started the app.

Three-quarters of all respondents lived in Türkiye itself. Most of the remaining quarter lived in the UK or the US, both of which are English-speaking countries. Despite the large number of Turkish residents in Germany, only three participants came from the Turkish diaspora in that country.

The number of monolingual Turkish speakers and the number of bi- and multilingual speakers was far more balanced, with 57% of the sample being monolingual Turkish speakers and 43% being bilinguals.

Of all participants who completed the survey, 16 (6%) reported a neurocognitive or a psychiatric diagnosis, but just 4 (25%) of these participants continued to the SVF application and provided accepted recordings. The diagnoses reported were depression (N = 1), hyperactivity and high IQ (N = 1), and mood disorder (N = 2).

Multilingualism. Detailed information on participants' linguistic background is provided in Table 6.7. Forty-five (43%) participants reported being multilingual and declared proficiency in at least one language other than Turkish. Three-quarters were bilingual (N = 33, 73%), and the remaining quarter was multilingual (N = 12, 27%). We defined a person's second language as the language with the highest overall language skills scores. The most frequent second language was English (N = 39, 87%), followed by Arabic (N = 3, 7%), Kurdish (N = 2, 4%), and German (N = 1, 2%). Other languages listed included Korean, Spanish, Japanese, Zaza, Macedonian, Swedish, Norwegian, and Slovak. The average age at which the second language was acquired was 13, ranging from 1 to 30 years. On average, individuals reported using languages other than their native language for approximately 40% of their lives. The average self-

rating on individual skills was between 3.5 and 4, which translates to being largely able to engage with complex texts and conversations but not at the level of full, near-native competence in both languages.

		Animals				**Veg&Fruits**				**Supermarket Items**			
		N = 45				**N = 45**				**N = 43**			
		Mean	SD	max	min	Mean	SD	max	min	Mean	SD	max	min
No. of Other Languages		1.37	0.71	4	1	1.4	0.71	4	1	1.39	0.72	4	1
Age of L2 acquisition		12.8	6.87	30	1	13.64	7.35	30	1	13.93	7.2	30	1
Percentage of L2 usage		38.86	24.28	95	3	39.64	25.41	95	0	41.02	25.62	95	3
Language ability	Overall	3.93	0.88	5	2	3.86	0.94	5	2	3.88	0.9	5	2
	Speaking Casual	3.84	1.06	5	1	3.77	1.1	5	1	3.76	1.06	5	1
	Listening Casual	3.82	0.98	5	1	3.77	1.04	5	1	3.76	0.99	5	1
	Speaking Formal	3.51	1.25	5	1	3.42	1.3	5	1	3.46	1.27	5	1
	Listening Formal	3.62	1.23	5	1	3.55	1.25	5	1	3.58	1.25	5	1
	Reading	3.64	1.2	5	1	3.62	1.24	5	1	3.67	1.24	5	1
	Writing	3.44	1.21	5	1	3.42	1.28	5	1	3.48	1.26	5	1
One foreign language(N)		33				32				31			
Multiple foreign language(N)		12				13				12			

Table 6.7: Language use and language skills reported by bi- and multilingual participants, including their self-rated language ability scores.

6.10 Discussion

In this chapter, we reported on the collection and preparation of a corpus of Turkish SVF data that will be made accessible to researchers wishing to develop and test automatic SVF analysis algorithms for Turkish. To the best of our knowledge, this is the first open-source SVF corpus in Turkish and the first time that Turkish SVF data has been collected online. In our systematic review (Chapter 4), we established that the most commonly used data collection techniques for the SVF test in the Turkish literature were paper and pencil assessment Akdemir (2021) and voice recording Tuncer (2012); Özdemir (2015), with the help of a clinical practitioner in both cases. Our corpus might even be the first SVF dataset collected online for any language. In our review of the SVF literature, we found only one study that had used a web-based and self-paced data collection method to collect a corpus of data for further analysis. However, the sequences collected were for the phonemic verbal fluency test (letter F) (Cho et al., 2021). Furthermore, Cho et al. (2021) recruited subjects (N = 76) from univer-

sity volunteers, not from the general population, and did not make the data collection tool freely available. In their data, Cho et al. (2021) encountered filler words (e.g. um, uh, and er) and non-verbal expressions (e.g. a laugh or cough) as well as the nouns from desired category and did not observe the phenomena recorded in our data, such as thinking aloud and speech from an assistant.

Online Data Collection: Online data collection is fully asynchronous, as was the case in this study. No practitioner is required to administer the test, and participants can take the test at their own pace in an environment where they feel relaxed. In terms of performance, our findings are comparable to the results from normative studies, but future work is needed to establish whether metrics derived from SVF sequences collected online are sufficiently similar to metrics derived from SVF sequences that are collected in person. Such validation is important if normative findings established using paper and pencil assessment are to be applied to findings established using online data. In particular, we found substantial amounts of **extra speech** (24% of all Animal sequences, 20% of all Fruit & Vegetables sequences, and 15% of all Supermarket Item sequences), which may affect word counts. When SVF sequences are obtained in person, the practitioner can use non-verbal and verbal cues to encourage the participant to continue, which was not an option in our app. As a result, participants may have thought aloud more or hesitated for longer than they might have done otherwise.

We encountered substantial **dropout** between survey completion and SVF data collection. In total, 170 participants quit at various stages of the study; most dropped out after the survey (N = 126, 48% of completed surveys). Dropout is a common problem in online studies. In the early days of web-based studies, Birnbaum (2004) found that web-based studies had a very high dropout rate because participants were able to quit easily without having to provide a reason. Reasons for drop-out may include lack of motivation, technical difficulties, or unexpected sensitive questions Bosnjak and Tuten (2001); Jun et al. (2017). The relative ease of completing the demographic survey may have contributed to the completion of the survey stage by those who subsequently dropped out. While some may have left the study due to perceiving the spoken data to be too sensitive and easily identifiable, technical problems are the most likely reason for drop outs. Since the system was built on two separate platforms and connected through user ID input, participants had to either note down their participant number after the survey and manually input it into the recording app or else copy it to clipboard and paste it into the text box in the app. Users may have forgotten to record their

ID, made a mistake when noting it down, or found the procedure too cumbersome. Two suggestions that could address this issue are worth considering for researchers developing a similar application or those interested in using our application. Firstly, a connection system could be established where the voice recording system automatically retrieves the ID number generated after the survey, but overcoming the challenge of linking different systems would be necessary. Additionally, merging the two systems under a single umbrella platform by creating an integrated all-in-one tool could reduce the number of drop-outs resulting from the transition between the two platforms. However, this might introduce privacy issues, such as the risk of data leakage if demographic data and voice recordings are stored in the same location, necessitating consideration of alternative measures to address confidentiality concerns.

According to Jun et al. (2017), the strongest motivation for participants to continue the online study is the desire to contribute to science. They reported that younger people are more likely to quit than older people, a tendency that we also found in our dataset. However, although older participants tend to be more persistent in continuing, the fact that our sample consisted predominantly of younger participants may have skewed the statistics, leading to higher numbers of drop outs among the younger demographic. Although participants start with high motivation, the idea of being directed solely by written instructions without admin assistance may have made participants feel uncomfortable, especially in the recording stage, where they may have felt uneasy about speaking in front of a screen. In this situation, even if the individual has the intention to contribute to science, they will likely abandon the system. Particularly, the presence of a countdown, difficulties in perceiving whether the time has started or not, or uncertainty about whether their voice has truly been recorded may emerge as factors negatively affecting motivation. However, this scenario appears to be more probable for a participant who has at least navigated to the voice recording section. A helpful suggestion to ease the anxiety of speaking in front of a screen is to provide instructions in an audible format, allowing participants to complete this section with an auto-generated voice guide. Finally, another factor that we suspect to be involved in the high drop-out rate is lack of a suitable financial incentive. Further work is needed to explore other reasons for dropout.

Issues related to item categories: Our data collection did not cover one category which is commonly used and reported in the Turkish literature on SVF—first names. To the best of our knowledge, Turkish is the only language where this particular cate-

gory is used for SVF sequences, which means that findings in this category are difficult to map onto similar data for other languages. While the Delis-Kaplan Test of Executive Function Delis et al. (2001a) also uses first names as part of their semantic verbal fluency tasks, the category is split into boys' names for the main version (animals and boys' names) and girls' names for the alternate version (clothing and girls' names). First name sequences are also potentially problematic for computational linguistic approaches that rely on large corpora to assess name similarity, because uncommon first names might be missing or underrepresented. Finally, sequences of uncommon first names may be used to identify a participant in case they list the names of all their family members.

In the Supermarket items category, we observed that participants use generic brand names when listing the products they can purchase from the market. For instance, common brands like 'Selpak' for tissues (like Kleenex in United Kingdom)or 'Halley' for biscuits (like Lotus in United Kingdom) were mentioned among the products. Although 'Selpak' is not a meaningful word in Turkish, the name 'Halley' comes from Halley's Comet, which might lead the algorithm to place this word in a vector space associated with astronomy-related terms rather than food items and potentially affect the results. A similar situation may occcur in English for 'Lotus'. It is a flower name, so the vector representation may be less likely to suggest a biscuit. We kept those brand names in our analysis step, since we had a limited number of examples and wanted to maintain the provided relationships. However, during data collection, providing instructions to participants to avoid using proper names and instead mention the product category (e.g. tea, biscuits, chocolate, tissues) could facilitate data analysis for researchers and potentially improve the performance of computational linguistics tools.

Limitations of the Corpus: Unlike normative studies of Turkish SVF, our data exhibits a substantial sampling bias. Participants were well-educated, younger, and either employed full time or students. Most of the members of the Turkish diaspora who took part in the survey were from the UK or the US. Other countries where a significant number of Turkish speakers live, such as Germany or Denmark, were substantially underrepresented. In terms of self-reported language proficiency, we note that most participants situate themselves at a level that is well below near-native competence. Further studies of bilingual speakers of Turkish should use more in-depth assessments of language ability and language dominance. Finally, participation in the

study required internet access and access to a device with a microphone. This is likely to disproportionately exclude people of low socioeconomic status, older people, and people with a lower level of education, especially within Türkiye (Aytuna and Çapraz, 2018).

6.11 Conclusion

In this chapter, we presented a unique corpus of Turkish SVF sequences that was collected fully online and will be distributed to researchers. This is essential for reproducible research on the computational analysis of Turkish SVF data. Further work is needed to reduce the dropout rate and achieve a dataset that can be considered normative. In the next chapter, we present the computational analysis tools we developed for Turkish and attempt to replicate relevant Turkish SVF findings from the literature, which were summarised in Chapter 4.

Chapter 7

Analysis of Turkish Semantic Verbal Fluency Data Collected Online

7.1 Overview

The aims of the study presented in this chapter were twofold. First, we analysed the dataset described in Chapter 6, establishing whether we were able to replicate the results for demographic differences reported in Chapter 4. Secondly, as we did for Spanish data in Chapter 5, we analysed the internal structure of Turkish SVF sequences using computational methods and evaluated the utility of those measures for the Turkish SVF dataset described in Chapter 6. Specifically, our research questions were:

1. Dataset Analysis:

    (a) What are the descriptive statistics of the dataset in terms of word count, perseverations, clustering, and switching?

    (b) Are there significant differences in any of those metrics between genders (male/female) and between monolingual and bi-/multilingual speakers?

2. Computational Analysis:

    (a) To what extent can computational methods replicate the internal structure of Turkish SVF Animal sequences, as established using the Troyer method?

    (b) What is the internal structure of the three types of SVF sequences collected in the Turkish SVF dataset, as established using the computational methods?

(c) If there are significant differences in cluster size or number of switches between demographic categories in the Turkish SVF dataset, can these differences be replicated when clusters and switches are determined using computational methods?

As in Chapter 5, we used two computational methods, bigram analysis and Word2vec vector-space analysis to address research questions above.

7.2 Methodology

7.2.1 Dataset

The data sample collected in the Turkish data collection study was used (for details, see Chapter 6). The demographic characteristics of the dataset are given in Table 6.6 in Chapter 6.

7.2.2 Manual Annotation: Troyer Method

In order to implement the Troyer Method (Section 2.4.2), the taxonomy needs to be adapted for the Turkish language, since it was published in English and includes a limited number of animal names. We initially translated the original animal groups and names into Turkish. Then, we expanded the predefined groups by adding new animal names that were not in the original list but were produced by the participants. This process was completed by the author, who is a native Turkish speaker. The expanded Turkish version of the Troyer taxonomy list can be found in Appendix A.10. Clusters and switches were created for each participant according to the guidelines indicated in Troyer et al. (1997), using the adapted Turkish version of the Troyer taxonomy that we had created. Figure 7.1 shows a sample SVF sequence produced by a participant, divided into clusters and switches.

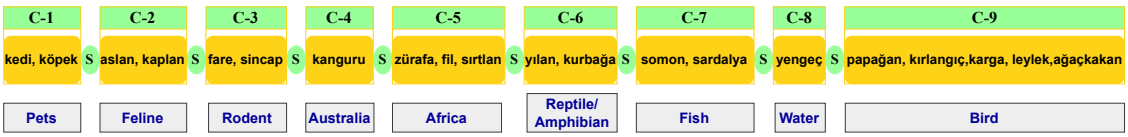


Figure 7.1: Troyer clusters and switches of sample Turkish SVF sequence

7.2.3 Computational Baseline: Bigram

As in Chapter 5, the bigram technique was used as a simple computational baseline. Details of the method and how to create clusters and switches are described in Section 3.2.

Overall, there were 2538 bigram tokens (sequences of two animal names) and 1593 bigram types (unique animal pairs) in the dataset. 'kedi-köpek' (cat–dog) was the most frequent animal pair, occurring 67 times, followed by 'aslan–kaplan' (lion–tiger) with 48 occurrences. Out of the total number of animal pairs, n = 1208 pairs (76%) were observed only once in the dataset, while n = 385 pairs (24%) were observed more than once.

7.2.4 Vector Space Model

We now describe the adaptation of the vector space model (Section 3.5) to Turkish, following a process similar to that used in Chapter 5. The Wikidump used was based on the Turkish Wikipedia as of 01 January 2023[1]. The dump included 436,483 articles with a size of approximately 2.8 GB.

The pre-processing steps were explained in Section 3.5.1.2. For stopword removal, we used the Turkish stopword list provided in NLTK (Bird and Loper, 2004) version 3.7. We tested two methods for removing affixes, lemmatisation and stemming. For lemmatisation, we used the the **Zeyrek** [2] morphological analyser and lemmatiser, which is a part of the **Zemberek-NLP** [3]; natural language processing library for Turkish (Akın and Akın, 2007). For stemming, we used the Turkish version of the Snowball (Porter, 2001) stemmer, **TurkishStemmer** [4] (Çilden, 2006) as implemented in NLTK version 3.7.

7.2.4.1 Model Creation

Again, we created a total of 12 models from shallow to deep using different hyperparameters: **Architecture:** CBOW and Skip-gram; **Window-Size:** 4 and 10; **Dimensions:** 300, 600, and 1000.

[1]Detailed information about the size and key features of the Turkish Wikipedia: `https://dumps.wikimedia.org/trwiki/`

[2]Zeyrek lemmatiser tool page: `https://zeyrek.readthedocs.io/en/latest/`

[3]Zemberek-NLP library documentation page: `https://github.com/ahmetaa/zemberek-nlp`

[4]Turkish Snowball stemmer algorithm page: `https://snowballstem.org/algorithms/turkish/stemmer.html`

The threshold values for each of the 12 models were determined based on a calculation of similarity scores between animal names in our dataset. They are shown with whisker plots in Figure 7.2. The red lines show the value of the 50th percentile, the lower line of each box indicates the value of the 25th percentile, and the upper line of each box indicates the value of the 75th percentile.

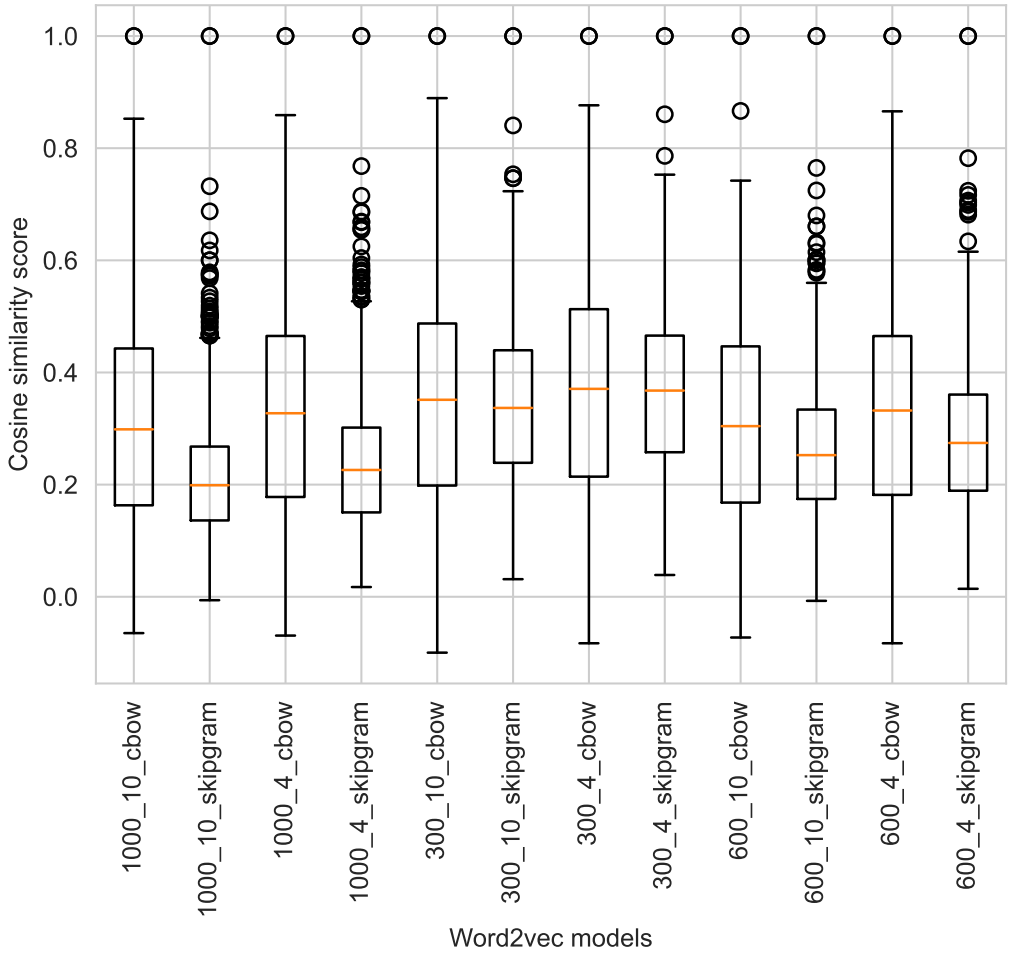


Figure 7.2: Threshold values of 12 Word2vec models through Turkish animal SVF dataset, named according to the following pattern: Dimension_WindowSize_Architecture. As an example, the threshold values for the 1000_10_cbow are as follows: 25th = 0.16, 50th = 0.29 and 75th = 0.44.

As in Chapter 5, we used Spearman's rank-order correlation coefficient to compare the number of switches and mean cluster size to the results from manual annotation. We also assessed the location of switches between clusters as the switch locations produced by one model may differ from those of another model.

7.3 Results

7.3.1 Baseline: Descriptive Statistics of Traditional Metrics

Table 7.1 shows the word count metrics for the dataset, while Table 7.2 provides descriptive statistics for perseverations. Category violations are reported in Table 6.5 in Chapter 6.

The descriptive statistics indicate that the Supermarket Items category is the lowest in word repetition and the highest in word production, making it the richest category in terms of word variety. Both clustering and switching are examined in detail for each category below.

		Total word count																	
		Animals						**Veg&Fruits**						**Supermarket**					
	Group Features	**N**	Mean	Mdn	Max	Min	Std	**N**	Mean	Mdn	Max	Min	Std	**N**	Mean	Mdn	Max	Min	Std
Age	18-29 age	18	29.83	28	50	17	8.75	19	25.68	25	36	17	4.26	16	27.56	27.5	41	17	6.95
	30-39 age	52	23.63	23	39	11	6.95	52	25.17	26.5	38	6	5.92	52	25.5	24.5	46	12	7.83
	40-49 age	19	26.68	26	51	13	10.02	18	27.17	27.5	44	15	7.88	17	28.53	25	43	17	7.69
	50-59 age	11	24.09	25	34	11	7.05	11	24.82	26	34	14	5.72	11	25.09	25	37	10	8.01
	60 age +	5	17.6	20	25	11	6.31	5	20	21	23	14	3.54	5	19	21	26	11	5.7
Gender	Male	51	26.27	26	51	11	7.64	51	25.18	25	44	6	6.68	50	27.2	24.5	46	12	8.28
	Female	53	24.06	24	50	11	8.59	53	25.68	27	36	14	5.17	50	24.9	25.5	37	10	7.03
	Non-binary	1	11	11	11	11	N/A	1	14	14	14	14	N/A	1	18	18	18	18	N/A
Education level	PhD	15	28.07	29	38	13	6.84	15	25.53	27	36	15	6.06	15	30.2	28	41	23	6.41
	Msc	29	24.28	24	37	11	6.55	29	26.48	27	41	16	5.48	30	26.03	25.5	46	12	8.6
	Bachelor degree	44	25.61	24	51	12	8.79	43	25.51	25	44	6	6.39	39	25.54	23	43	13	7.39
	Associate degree	3	18.67	20	25	11	7.09	3	20.67	20	23	19	2.08	3	21.33	22	26	16	5.03
	High school (9-12 years)	7	25.29	25	50	11	12.45	7	24	24	31	14	5.69	6	27.17	26.5	37	18	6.49
	Secondary School(6-8 years)	0	N/A	N/A	N/A	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A
	Primary school (1-5 years)	7	20.14	20	33	11	8.34	8	22.62	21.5	34	14	6.63	8	20.75	21.5	31	10	7.11
Occupation Status	Retired	4	17.5	17	25	11	6.24	4	19.75	21	23	14	4.27	4	19	20	26	10	6.83
	Homemaker	9	21	20	33	11	7.33	10	24.4	24.5	34	14	6.62	10	26	27	37	11	7.04
	Self-employment	3	24	26	29	17	6.24	2	28.5	28.5	30	27	2.12	2	24	24	25	23	1.41
	Full time employee	57	25.89	25	45	11	8.2	55	25.73	27	44	6	6.38	51	27.61	27	43	14	7.75
	Part time employee	3	24.67	24	29	21	4.04	3	21.67	21	25	19	3.06	3	26	25	32	21	5.57
	Student	20	24.65	24.5	50	15	8.12	23	24.87	24	38	16	5.47	22	24.64	24	46	12	7.81
	Unemployed	2	22	22	27	17	7.07	2	25.5	25.5	27	24	2.12	2	24.5	24.5	25	24	0.71
	Other	7	29.71	24	51	20	11.27	6	29.33	28	35	25	4.97	7	23.14	22	36	12	10.04
Place of residence	Germany	1	34	34	34	34	N/A	1	24	24	24	24	N/A	1	41	41	41	41	N/A
	USA	4	20.25	19.5	29	13	6.7	3	25.33	27	30	19	5.69	3	21.67	25	26	14	6.66
	United Kingdom	24	24	24.5	37	15	5.85	26	24.5	25	38	15	5.89	26	26.08	24.5	46	12	8.09
	Other	0	N/A	N/A	N/A	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	1	12	12	12	12	N/A
	Türkiye	76	25.46	25	51	11	8.91	75	25.63	26	44	6	6.15	70	26.1	25	43	10	7.4
Lang.	Monolingual	60	24.55	24.5	50	11	8.88	60	25.02	25	41	6	6.24	58	26.05	25.5	41	10	7.37
	Bi/Multilingual	45	25.62	26	51	13	7.39	45	25.73	26	44	15	5.73	43	25.86	25	46	12	8.29
	Total	105	25.04	25	51	11	8.26	105	25.32	25	44	6	6.01	101	25.97	25	46	10	7.73

Table 7.1: Total word counts for the three semantic categories by demographic features

Total word count: This includes all words belonging to the specified category, including perseverations. The maximum word counts observed were 51 in Animals, 46 in Supermarket Items, and 44 in Fruits & Vegetables. The minimum word counts were

as follows: Fruits & Veg (Min = 6) < Supermarket Items (Min = 10) < Animals (Min = 11). For Animals, the minimum of 11 words was the same across many demographic groups, lowering the average numbers for the Animal category. The mean number of words produced was between 25 and 26 for all three categories, with Supermarket Items (Mean = 25.97) > Fruits & Veg (Mean = 25.32) > Animals (Mean = 25.04).

		Perseverations																	
		Animals						**Veg&Fruits**						**Supermarket**					
	Group Features	**N**	Mean	Mdn	Max	Min	Std	**N**	Mean	Mdn	Max	Min	Std	**N**	Mean	Mdn	Max	Min	Std
Age	18-29 age	18	1.28	0	10	0	3	19	1.68	1	8	0	2	16	1.12	1	4	0	1.36
	30-39 age	52	0.79	0	5	0	1	52	1.33	1	9	0	2	52	0.73	0	5	0	1.21
	40-49 age	19	2.21	1	11	0	3	18	3.28	2	14	0	4	17	1.47	1	6	0	2
	50-59 age	11	1.27	1	6	0	2	11	1	0	5	0	2	11	0.91	1	4	0	1.22
	60 age +	5	1.6	1	3	0	1	5	1.6	1	3	0	1	5	0.2	0	1	0	0.45
Gender	Male	51	1.37	1	11	0	2	51	2	1	14	0	3	50	1.1	0	6	0	1.68
	Female	53	1.08	0	10	0	2	53	1.45	1	8	0	2	50	0.72	0	4	0	1.01
	Non-binary	1	1	1	1	1	N/A	1	0	0	0	0	N/A	1	1	1	1	1	N/A
Education level	PhD	15	0.73	0	5	0	1	15	1.13	1	4	0	1	15	1.4	1	5	0	1.64
	Msc	29	0.69	0	4	0	1	29	1.72	1	14	0	3	30	0.87	0	6	0	1.36
	Bachelor degree	44	1.7	1	11	0	3	43	1.81	1	9	0	2	39	0.87	0	6	0	1.44
	Associate degree	3	1.33	1	3	0	2	3	2.67	3	3	2	1	3	0.67	1	1	0	0.58
	High school (9-12 years)	7	1.71	1	6	0	2	7	2.29	2	8	0	3	6	1	0.5	4	0	1.55
	Secondary School(6-8 years)	0	N/A	N/A	N/A	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A
	Primary school (1-5 years)	7	0.86	1	3	0	1	8	1.25	0.5	5	0	2	8	0.38	0	2	0	0.74
Occupation Status	Retired	4	0.75	1	1	0	1	4	2.5	2.5	5	0	2	4	0.75	1	1	0	0.5
	Homemaker	9	0.89	0	3	0	1	10	1.2	0.5	5	0	2	10	0.9	0	4	0	1.37
	Self-employment	3	1.4	0	11	0	2	2	1.5	1.5	3	0	2	2	0.5	0.5	1	0	0.71
	Full time employee	57	0.33	0	1	0	1	55	1.85	1	14	0	2	51	1.12	0	6	0	1.67
	Part time employee	3	1	1	2	0	1	3	1	1	2	0	1	3	0.33	0	1	0	0.58
	Student	20	0.65	0	4	0	1	23	1.43	1	9	0	2	22	0.5	0	2	0	0.74
	Unemployed	2	3	3	6	0	4	2	1.5	1.5	2	1	1	2	0.5	0.5	1	0	0.71
	Other	7	2	2	6	0	2	6	2.17	2	4	0	2	7	1.29	1	4	0	1.5
Place of residence	Germany	1	0	0	0	0	N/A	1	1	1	1	1	N/A	1	4	4	4	4	N/A
	USA	4	0.5	0	2	0	1	3	2.33	2	5	0	3	3	0.33	0	1	0	0.58
	United Kingdom	24	0.58	0	2	0	1	26	1.15	1	9	0	2	26	0.62	0	5	0	1.1
	Other	0	N/A	N/A	N/A	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A	1	1	1	1	1	N/A
	Türkiye	76	1.47	1	11	0	2	75	1.88	1	14	0	2	70	1	0	6	0	1.46
Lang.	Monolingual	60	1.47	0.5	11	0	2	60	1.63	1	14	0	2	58	1.09	1	6	0	1.47
	Bi/Multilingual	45	0.89	0	6	0	1	45	1.8	1	9	0	2	43	0.67	0	5	0	1.25
	Total	105	1.22	0	11	0	1.98	105	1.7	1	14	0	2.25	101	0.91	0	6	0	1.38

Table 7.2: Perseverations for the three semantic categories by demographic features

Perseverations: Whereas some sequences contained no perseverations, the maximum number of perseverations recorded was 14 for Fruits & Veg, 11 for Animals and 6 for Supermarket Items. On average, the number of perseverations in each sequence varied between 1 and 2 (Fruits & Veg (Mean = 1.7, Max = 14) > Animals (Mean = 1.22, Max = 11) > Supermarket Items(Mean = 0.91, Max = 6))

Clustering and Switching: Clustering and switching were only investigated for the Animal category. Table 7.3 shows the results. Although the number of switches varied between 3 and 38, participants tended to make few switches. The mean cluster size

showed a distribution similar to that of switch numbers, and clusters with few elements were more common. The mean cluster size ranged between 3.78 and 1.07.

								T Test		
		N	**Mean**	**Mdn**	**Max**	**Min**	**Std**	***t*-value**	***p*-value**	η^2
Number of switches										
	Gender(M/F)	51/53	11.71/12.53	11.0/12.0	23/38	3/3	3.61/5.14	-0.932	0.353	
	Language(Mono/Bi)	60/45	11.83/12.4	11.0/13.0	38/23	3/6	5.06/3.54	-0.637	0.526	
	All participants	105	12.08	12	38	3				
Mean Cluster size										
	Gender(M/F)	51/53	2.11/1.84	2.07/1.67	3.78/3.43	1.38/1.07	0.46/0.47	2.965	0.004**	0.582
	Language(Mono/Bi)	60/45	1.98/1.95	1.92/1.88	3.78/3.43	1.07/1.38	0.53/0.43	0.299	0.765	
	All participants	105	1.97	1.91	3.78	1.07				

Table 7.3: Switch and cluster components derived from the Troyer method, sorted three ways: all participants, by gender (male/female), by language (monolingual/bi-multilingual).

7.3.2 Group Differences

Since our dataset was skewed towards younger and well-educated participants, it was not possible to meaningfully replicate differences with respect to age and socioeconomic status that have been documented in the normative literature on Turkish SVF. Therefore, we only compared participants' SVF performance across genders (male versus female) and language statuses (mono- versus bi-/multilingual). T-Tests were used to establish significance. The results are summarised in Table 7.4. We found no significant difference in either demographic grouping in any semantic category, which is in line with previously reported findings for gender.

According to the clustering and switching features, showed in Table 7.3, we found no difference between males and females in the number of switches but did obtain a significant difference in the mean cluster size ($t_{(104)} = 2.965$, $p<0.01$). The effect size was $\eta^2 = 0.582$, so males tend to create larger clusters and are approximately 0.6 standard deviations above females in mean cluster size. In terms of language status, there is no difference between monolinguals and bilinguals in either switching or clustering.

									T Test		
		Category	N	Mean	Mdn	Max	Min	Std	t-value	p-value	η^2
Total words											
	Gender(M/F)	Animals	51/53	26.27/24.06	26/24	51/50	11/11	7.57/8.51	1.389	0.168	0.273
		Veg\&Fruits	51/53	25.18/25.68	25/27	44/36	6/14	6.61/5.12	-0.43	0.668	-0.084
		Supermarket	50/50	27.2/24.9	24.5/25.5	46/37	12/10	8.2/6.96	1.497	0.138	0.299
	Language(Bi/Mono)	Animals	45/60	25.62/24.55	26/24.5	51/50	13/11	7.3/8.81	0.657	0.513	0.13
		Veg\&Fruits	45/60	25.73/25.02	26/25	44/41	15/6	5.67/6.19	0.602	0.548	0.119
		Supermarket	43/58	25.86/26.05	25/25.5	46/41	12/10	8.19/7.3	-0.122	0.903	-0.025
Perseverations											
	Gender(M/F)	Animals	51/53	1.37/1.08	1/0	11/10	0/0	2.2/1.75	0.757	0.45	0.149
		Veg\&Fruits	51/53	2/1.45	1/1	14/8	0/0	2.68/1.7	1.236	0.219	0.243
		Supermarket	50/50	1.1/0.72	0/0	6/4	0/0	1.66/1	1.369	0.174	0.274
	Language(Bi/Mono)	Animals	45/60	0.89/1.47	0/0.5	6/11	0/1	1.29/2.33	-1.484	0.141	-0.293
		Veg\&Fruits	45/60	1.8/1.63	1/1	9/14	0/1	2.01/2.41	0.372	0.71	0.074
		Supermarket	43/58	0.67/1.09	0/1	5/6	0/1	1.23/1.45	-1.484	0.141	-0.299

Table 7.4: Total word count and perseveration T-test statistics in terms of gender and language status for the three categories separately.

7.3.3 Computational Methods versus Manual Annotation

7.3.3.1 Bigram

According to the Troyer method, the mean switch number was 12.08 in the Animal category. The bigram method gave similar results, with a mean switch number of 11.53 (see Table 7.5 for descriptive statistics of clustering and switching features). When we look at the cluster sizes, The cluster sizes obtained with the Troyer method vary between 1.07 and 3.78, and those calculated using the bigram method range from a minimum of 1.2 to a maximum of 5.5. Additionally, the mean cluster size in the bigram method was 0.22 more than that in the Troyer method, suggesting that the bigram method creates relatively bigger clusters and fewer switches. In other categories, the mean switch numbers obtained using the bigram method are 15.81 for supermarket and 9.44 for vegetables and fruits. Since the bigram method depends on the frequency and only frequency=1 generates a switch, participants produced more diverse nouns in the Supermarket Items category compared to the other two categories. Name pairs were observed less frequently and this likely resulted in more switches. In the category of Fruits and Vegetables, the bigram method yielded the lowest mean score in switch numbers (9.44) and the biggest mean cluster size (2.61). This means that people produced words that were commonly said by another participants.

					Animal Category								Supermarket Category								Vegetables and Fruits Category							
Hyperparameters					Number of switches				Mean Cluster size				Number of switches				Mean Cluster size				Number of switches				Mean Cluster size			
Model	Threshold	d	w	f	Mean	Mdn	max	min	Mean	Mdn	max	min	Mean	Mdn	max	min	Mean	Mdn	max	min	Mean	Mdn	max	min	Mean	Mdn	max	min
Word2vec	0.75	1000	10	cbow	17.01	16	42	6	1.4	1.36	2.2	1.1	17.55	17	32	5	1.41	1.38	2.16	1	16.6	17	38	2	1.48	1.42	2.6	1.1
		1000	**10**	**skipgram**	15.69	15	47	5	1.52	1.5	2.3	1	17.82	18	35	8	1.38	1.33	2.33	1	16.3	17	32	1	1.51	1.44	3	1.1
		1000	4	cbow	17.31	17	44	6	1.38	1.33	2.2	1.1	17.79	17	32	5	1.38	1.33	1.93	1	17	17	36	2	1.43	1.4	2.6	1.1
		1000	4	skipgram	16.92	16	43	6	1.41	1.38	2.2	1.1	18.01	18	36	7	1.37	1.35	2	1	17.6	18	34	2	1.39	1.36	2.1	1.1
		600	10	cbow	16.9	16	42	5	1.42	1.38	2.4	1.1	17.59	17	31	5	1.4	1.33	2.16	1	16.5	17	37	2	1.48	1.43	2.6	1.1
		600	**10**	**skipgram**	15.65	15	46	5	1.53	1.5	2.3	1.1	17.66	17	36	6	1.4	1.33	2.33	1	15.9	16	32	1	1.56	1.47	3	1.1
		600	4	cbow	17.26	17	46	6	1.4	1.33	2.9	1.1	17.86	17	31	5	1.38	1.33	1.95	1	17	17	38	2	1.44	1.4	2.6	1.1
		600	4	skipgram	16.59	16	45	6	1.44	1.43	2.3	1.1	17.96	18	35	7	1.38	1.35	2.62	1	17.3	18	33	2	1.41	1.4	2.1	1.1
		300	10	cbow	16.9	16	42	6	1.42	1.38	2.2	1.1	17.61	17	32	5	1.4	1.38	2.38	1	17	17	37	2	1.44	1.4	2.6	1.1
		300	10	skipgram	15.74	15	44	6	1.52	1.5	2.1	1.1	17.8	18	35	6	1.39	1.36	2	1	15.7	16	35	2	1.56	1.47	2.6	1.1
		300	4	cbow	17.17	17	43	6	1.39	1.34	2.2	1.1	17.75	17	31	5	1.39	1.35	1.95	1	16.9	16	37	3	1.43	1.4	2.6	1.1
		300	4	skipgram	16.62	16	42	6	1.44	1.41	2.2	1.1	17.97	18	36	7	1.37	1.36	2.33	1	17	17	37	2	1.44	1.4	2.3	1
	0.50	1000	10	cbow	10.9	11	29	3	2.2	2.08	4	1.5	11.59	12	23	3	2.18	2	5	1.21	10.3	10	24	0	2.5	2.23	10	1.3
		1000	10	skipgram	10.02	9	31	3	2.39	2.33	5	1.3	11.69	11	22	4	2.12	2	4.33	1.09	10.1	10	22	0	2.47	2.25	6	1.3
		1000	4	cbow	11.15	10	28	3	2.17	2	4.1	1.4	11.63	12	23	3	2.16	2	4.75	1.17	11.1	11	23	1	2.27	2.07	7.8	1.3
		1000	4	skipgram	10.79	10	31	3	2.23	2.2	4	1.2	11.91	11	24	4	2.09	1.91	4.2	1.09	11.1	11	25	0	2.26	2.08	6	1.2
		600	10	cbow	11.03	10	29	3	2.2	2.07	4	1.4	11.5	11	24	2	2.22	2.08	7	1.13	10.5	10	24	0	2.41	2.18	7.8	1.3
		600	10	skipgram	10.27	10	32	3	2.33	2.22	4	1.2	11.53	11	23	4	2.14	2	4.44	1.09	10.7	10	23	0	2.36	2.1	6.3	1.3
		600	4	cbow	10.98	10	29	3	2.2	2.07	4.1	1.4	11.7	11	24	3	2.15	2.07	4.75	1.09	11.2	11	23	1	2.27	2.08	7.8	1.3
		600	4	skipgram	10.69	10	31	3	2.24	2.22	4	1.3	11.81	11	23	4	2.1	1.94	4.33	1.09	10.8	10	23	0	2.3	2.18	6	1.2
		300	10	cbow	10.71	10	28	3	2.28	2	5	1.4	11.37	11	22	2	2.23	2.07	7	1.15	10.3	10	24	0	2.49	2.25	10	1.3
		300	10	skipgram	10.37	10	31	4	2.29	2.18	4	1.2	11.74	12	23	4	2.1	2	4.44	1.15	10.2	10	20	0	2.48	2.25	6.2	1.4
		300	4	cbow	11.1	11	30	3	2.16	2	4	1.5	11.66	11	24	3	2.16	2	5.17	1.09	11.2	11	23	1	2.26	2.07	7.8	1.3
		300	4	skipgram	10.65	10	31	3	2.24	2.22	4	1.4	11.73	11	25	4	2.1	2	4.2	1.2	10.5	10	25	0	2.41	2.11	6	1.3
	0.25	1000	10	cbow	4.8	4	17	0	5.22	4.25	21	2.1	6.02	6	14	0	4.61	3.71	22	1.67	4.5	4	14	0	5.84	4.5	26	2
		1000	10	skipgram	4.7	4	20	0	5.25	4.29	34	1.9	6.07	6	14	0	4.48	3.67	21	1.58	4.35	5	13	0	6.08	4.6	27	1.8
		1000	4	cbow	4.78	4	18	0	5.29	4.22	21	2.1	6.03	6	14	0	4.51	3.75	21	1.58	4.9	5	17	0	5.33	4.14	26	1.8
		1000	4	skipgram	4.85	4	17	0	5.24	4.33	34	1.8	6.11	6	14	0	4.4	3.5	13	1.46	4.5	5	12	0	5.96	4.5	30	1.8
		600	10	cbow	4.76	4	18	0	5.2	4.67	21	2.1	5.79	6	15	0	4.78	3.83	22	1.88	4.8	5	16	0	5.61	4.2	26	1.8
		600	10	skipgram	4.77	4	19	0	4.91	4.33	21	2	6.09	6	12	1	4.27	3.62	11	1.58	4.38	4	12	0	6.08	4.67	27	1.8
		600	4	cbow	4.72	4	19	0	5.41	4.57	22	2.1	6	6	14	0	4.56	3.7	22	1.73	4.87	5	16	0	5.39	4.2	26	1.8
		600	4	skipgram	4.84	4	19	0	5.28	4.33	34	2	6.17	6	14	0	4.39	3.62	22	1.58	4.53	4	12	0	5.89	4.5	27	1.8
		300	10	cbow	4.81	4	19	0	5.29	4.33	22	2.2	5.77	6	14	0	4.87	3.83	26	1.67	4.44	4	15	0	5.74	4.83	26	2
		300	10	skipgram	4.96	4	19	1	4.65	4.33	11	1.9	5.96	6	12	0	4.51	3.7	21	1.58	4.85	5	13	0	5.65	4.25	27	1.6
		300	4	cbow	4.79	5	18	0	5.35	4.25	21	2.1	6.01	6	15	0	4.49	3.78	22	1.67	4.91	5	15	0	5.32	4.2	26	1.8
		300	4	skipgram	4.76	4	18	0	5.08	4.33	21	2	6.08	6	13	0	4.6	3.88	22	1.58	4.49	4	11	0	5.74	4.5	27	1.8
Bigram					11.53	11	36	1	2.19	2.05	5.5	1.2	15.81	15	30	4	1.59	1.56	2.83	1.06	9.44	9	36	2	2.61	2.45	5.3	1.2

Table 7.5: Switching and clustering components derived from Bigram and 36 word2vec models representing the combinations of 12 hyperparameters and 3 thresholds with Snowball stemmer. The best performing models are bolded. Full descriptive statistics of the other morphological analyser, Zeyrek lemmatizer, are given in Appendix A.11

7.3.3.2 Determining the Best Vector Space Model

As can be seen from the Spearman's correlation results in Table 7.6, all models were highly correlated with manually obtained switch numbers. The best performing model with snowball stemmer was 600_10_skipgram, followed by the models 600_4_skipgram, 300_10_skipgram, and 1000_10_skipgram. We also found significant but weaker correlations between predicted and manually annotated cluster size.

The best performing models with respect to switch location (see Table 7.7) were 1000_10_skipgram (F1 = 0.738) and 600_10_skipgram (F1 = 0.734) with snowball stemmer. While recall was 0.84 for both models, precision was lower at 0.65. This suggests that vector space models tend to insert additional switches.

Figure 7.3 shows a two-dimensional representation of the vector space created by the model 1000_10_skipgram, a model that performs well for switch location, number of switches, and mean cluster size. We used principal component analysis (PCA) for dimensionality reduction. The three words 'karga', 'kırlangıç', and 'leylek' (crow-swallow-stork) are close semantically, while the pair 'kedi-köpek' (cat-dog) is far from 'kurbağa' (frog) due to the semantic dissimilarity of these words.

To understand the role of the different thresholds, we created Figure 7.4 using the mean columns from Table 7.5. In both figures, there are three line patterns, solid, dashed, and dotted, which represent the 0.75, 0.50, and 0.25 thresholds, respectively. Red, green, and blue indicate, in order, Animal, Supermarket Items, and Fruits & Vegetables. From the figure, it is clear that as cluster size increases, the number of switches decreases. For the 0.25 threshold, the models could not create any switches in some participants' sequences (the minimum number of switches equalled 0). If there is no switch in a sequence, it has a single large cluster consisting of all animal names produced by participant. The cluster size is increased at low threshold values because a relatively small similarity score is sufficient to assign two words to the same cluster. In that case, the model tends to group even unrelated words and clusters grow. With the 0.75 threshold, we see smaller clusters and more switches compared to the 0.25 and 0.50 threshold values.

7.3.3.3 Vector Space Model versus Bigram

The correlation between the number of switches obtained through manual annotation and the bigram method was ρ=0.87 (see Table 7.6). However, the cluster size correlation between the bigram and manual methods was very low. Switch location was

					Spearman Correlation								
					Switch Number					Mean Cluster Size			
	model hyperparameters				Zeyrek Lemma		Snowball Stem			Zeyrek Lemma		Snowball Stem		
	Threshold	d	w	f	p value	rho	p value	rho		p value	rho	p value	rho	
Word2vec	**0.75**	1000	10	cbow	1.17E-15	0.681867	2.34E-18	0.724733		0.006078	0.266083	1.21E-05	0.412774	
		1000	**10**	**skipgram**	1.91E-19	0.740016	2.69E-20	0.751258	*	0.000001	0.450159	1.16E-05	0.413596	*
		1000	4	cbow	2.96E-16	0.692038	1.10E-17	0.714752		0.007106	0.261245	1.21E-04	0.366446	
		1000	4	skipgram	4.30E-18	0.720856	4.45E-19	0.734959		0.000005	0.428684	2.20E-06	0.443172	
		600	10	cbow	1.03E-16	0.699584	6.03E-17	0.703277		0.001266	0.310486	3.01E-04	0.345947	
		600	**10**	**skipgram**	1.02E-18	0.72993	1.22E-21	0.767854	*	0.000023	0.40085	1.02E-07	0.491452	*
		600	4	cbow	6.44E-16	0.686334	8.89E-17	0.700579		0.00763	0.259015	2.11E-04	0.354101	
		600	4	skipgram	2.07E-16	0.694613	2.87E-21	0.763395	*	0.001767	0.301612	2.97E-08	0.508928	*
		300	10	cbow	1.91E-16	0.695164	3.49E-16	0.690838		0.002604	0.290935	2.98E-03	0.287153	
		300	10	skipgram	4.24E-18	0.720953	1.17E-20	0.755857	*	0.000257	0.349549	2.72E-06	0.439519	*
		300	4	cbow	5.44E-16	0.687588	6.32E-18	0.718376		0.030541	0.211229	1.46E-03	0.30672	
		300	4	skipgram	3.20E-16	0.691471	2.32E-18	0.724793		0.00208	0.297171	2.77E-04	0.347884	
	0.50	1000	10	cbow	2.99E-11	0.591729	5.44E-14	0.651186		0.028589	0.213719	2.20E-02	0.223436	
		1000	10	skipgram	7.89E-13	0.627504	4.17E-15	0.67212		0.002904	0.287863	9.98E-05	0.370555	
		1000	4	cbow	8.87E-11	0.580089	1.76E-12	0.62		0.430116	0.077812	6.66E-02	0.179699	
		1000	4	skipgram	5.53E-11	0.585203	2.91E-14	0.656449		0.001417	0.307517	1.67E-02	0.23302	
		600	10	cbow	1.31E-11	0.600253	2.02E-13	0.639834		0.021357	0.224427	9.03E-02	0.166141	
		600	10	skipgram	1.49E-11	0.598945	6.11E-16	0.686721		0.000174	0.358398	6.90E-06	0.423113	
		600	4	cbow	3.51E-12	0.613344	2.38E-12	0.61709		0.389528	0.084838	9.96E-02	0.161576	
		600	4	skipgram	1.49E-10	0.574399	7.73E-14	0.648181		0.03285	0.208451	3.02E-03	0.286716	
		300	10	cbow	8.95E-12	0.604095	5.39E-12	0.609139		0.142525	0.144088	2.89E-01	0.104444	
		300	10	skipgram	1.10E-10	0.577711	2.05E-14	0.659327		0.025995	0.217262	1.00E-04	0.370442	
		300	4	cbow	2.12E-11	0.59531	2.21E-13	0.639043		0.814966	0.023112	2.12E-01	0.122736	
		300	4	skipgram	1.51E-11	0.5988	2.80E-15	0.675222		0.015514	0.235666	1.35E-03	0.308822	
	0.25	1000	10	cbow	5.09E-05	0.384668	2.21E-08	0.512955		0.981527	-0.00229	7.93E-02	0.172017	
		1000	10	skipgram	2.76E-06	0.4393	4.46E-08	0.503242		0.471541	0.071024	1.81E-02	0.230232	
		1000	4	cbow	5.60E-05	0.382699	4.19E-08	0.504116		0.803755	-0.02454	3.97E-02	0.201047	
		1000	4	skipgram	2.25E-04	0.352659	2.51E-07	0.478025		0.322401	0.097506	2.64E-02	0.216621	
		600	10	cbow	3.87E-05	0.390244	9.69E-09	0.523968		0.823197	0.022067	4.92E-02	0.192436	
		600	10	skipgram	1.56E-07	0.485194	2.95E-10	0.56668		0.552074	0.058685	1.31E-03	0.309635	
		600	4	cbow	4.88E-05	0.385533	1.09E-07	0.490432		0.744661	-0.03216	1.62E-01	0.137318	
		600	4	skipgram	5.45E-06	0.427339	4.89E-08	0.501972		0.271328	0.10833	8.04E-02	0.171402	
		300	10	cbow	1.45E-05	0.409319	1.15E-08	0.521707		0.9464	0.00664	1.73E-02	0.231802	
		300	10	skipgram	1.68E-06	0.447662	1.44E-10	0.574793		0.915506	0.010479	4.50E-03	0.275147	
		300	4	cbow	2.36E-05	0.400043	1.15E-08	0.521735		0.654027	-0.04425	3.30E-01	0.095945	
		300	4	skipgram	2.51E-05	0.398778	1.33E-09	0.548982		0.65271	-0.04443	3.72E-02	0.203602	
					p value	rho				p value	rho			
Bigram					8.13E-35	0.878478				0.4273	0.07827			

Table 7.6: Table compares 36 Word2vec models and the bigram method with the Troyer method according to Spearman's correlation coefficient in terms of number of switches and mean cluster size. The comparison includes only animal category. A green colour gradient was used to emphasise how scores were changing.

					Switch Location						
	model hyperparameters				**Zeyrek Lemma**			**Snowball Stem**			
	Threshold	*d*	w	f	precision	recall	F1 score	precision	recall	F1 score	
Word2vec	**0.75**	1000	10	cbow	0.617	0.856	0.717	0.601	0.848	0.703	
		1000	**10**	**skipgram**	0.654	0.871	0.747	0.653	0.849	0.738	*
		1000	4	cbow	0.598	0.84	0.699	0.578	0.829	0.681	
		1000	4	skipgram	0.628	0.858	0.725	0.601	0.843	0.702	
		600	10	cbow	0.628	0.86	0.726	0.606	0.848	0.707	
		600	**10**	**skipgram**	0.645	0.864	0.739	0.65	0.843	0.734	*
		600	4	cbow	0.579	0.837	0.684	0.581	0.83	0.684	
		600	4	skipgram	0.626	0.848	0.72	0.612	0.841	0.708	
		300	10	cbow	0.626	0.85	0.721	0.599	0.84	0.699	
		300	10	skipgram	0.636	0.848	0.727	0.641	0.837	0.726	*
		300	4	cbow	0.588	0.832	0.689	0.576	0.82	0.677	
		300	4	skipgram	0.615	0.83	0.707	0.598	0.824	0.693	
	0.50	1000	10	cbow	0.623	0.578	0.6	0.625	0.565	0.593	
		1000	10	skipgram	0.653	0.577	0.613	0.682	0.566	0.619	
		1000	4	cbow	0.585	0.548	0.566	0.605	0.559	0.581	
		1000	4	skipgram	0.63	0.577	0.602	0.626	0.56	0.591	
		600	10	cbow	0.635	0.572	0.602	0.622	0.568	0.594	
		600	10	skipgram	0.655	0.577	0.614	0.686	0.583	0.63	
		600	4	cbow	0.591	0.549	0.569	0.604	0.549	0.575	
		600	4	skipgram	0.635	0.569	0.6	0.628	0.556	0.59	
		300	10	cbow	0.635	0.575	0.604	0.625	0.555	0.588	
		300	10	skipgram	0.642	0.569	0.603	0.656	0.564	0.607	
		300	4	cbow	0.581	0.545	0.562	0.586	0.539	0.562	
		300	4	skipgram	0.619	0.557	0.586	0.624	0.551	0.585	
	0.25	1000	10	cbow	0.581	0.273	0.371	0.698	0.278	0.398	
		1000	10	skipgram	0.639	0.282	0.391	0.746	0.29	0.418	
		1000	4	cbow	0.555	0.26	0.354	0.689	0.273	0.391	
		1000	4	skipgram	0.591	0.275	0.375	0.692	0.278	0.397	
		600	10	cbow	0.58	0.273	0.371	0.714	0.282	0.404	
		600	10	skipgram	0.617	0.282	0.387	0.743	0.294	0.421	
		600	4	cbow	0.555	0.259	0.353	0.677	0.265	0.381	
		600	4	skipgram	0.574	0.271	0.368	0.699	0.28	0.4	
		300	10	cbow	0.571	0.264	0.361	0.707	0.282	0.403	
		300	10	skipgram	0.595	0.268	0.37	0.691	0.284	0.403	
		300	4	cbow	0.553	0.257	0.351	0.67	0.266	0.381	
		300	4	skipgram	0.547	0.264	0.356	0.696	0.275	0.394	
Bigram					0.683	0.652	0.667				

Table 7.7: Table compares the bigram method and 36 Word2vec models with the Troyer method according to the switch locations. This comparison includes only the Animal category. A green colour gradient was used to emphasise how the scores were changing.

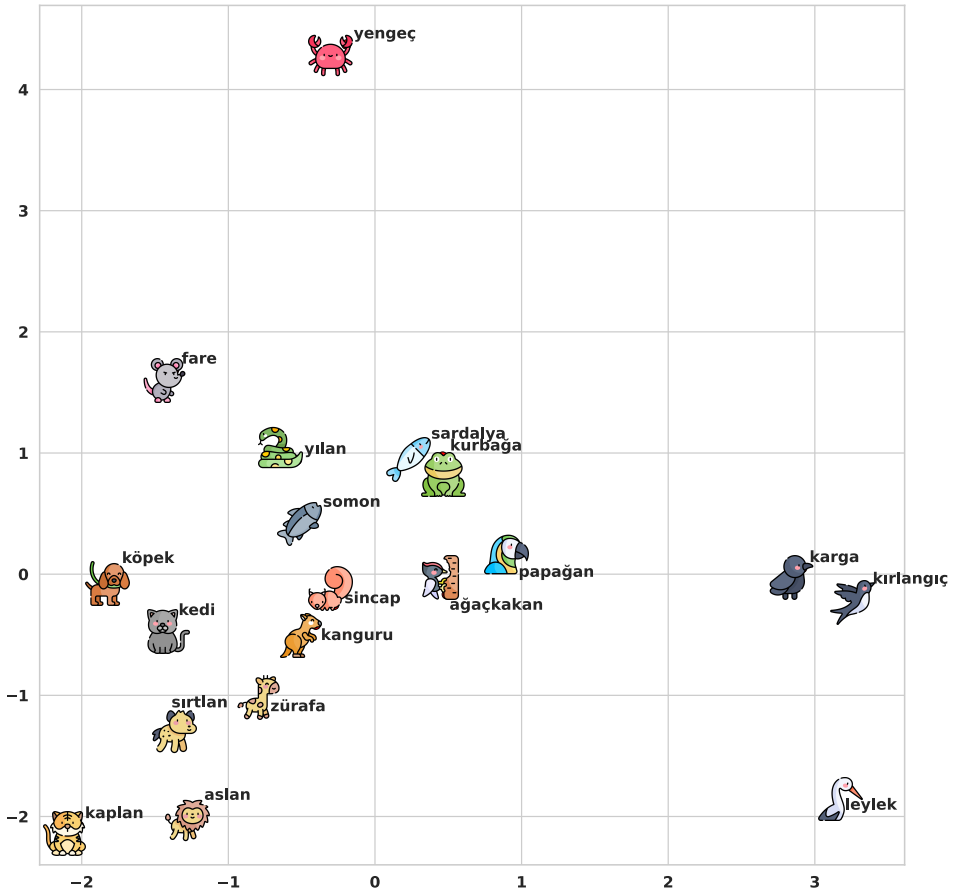


Figure 7.3: Two-dimensional PCA visualisation of the 1000_dimensional 10_window sized Skip-gram architectural vectors of Turkish animals. The projection exemplifies how the model captures the relationship between animals.

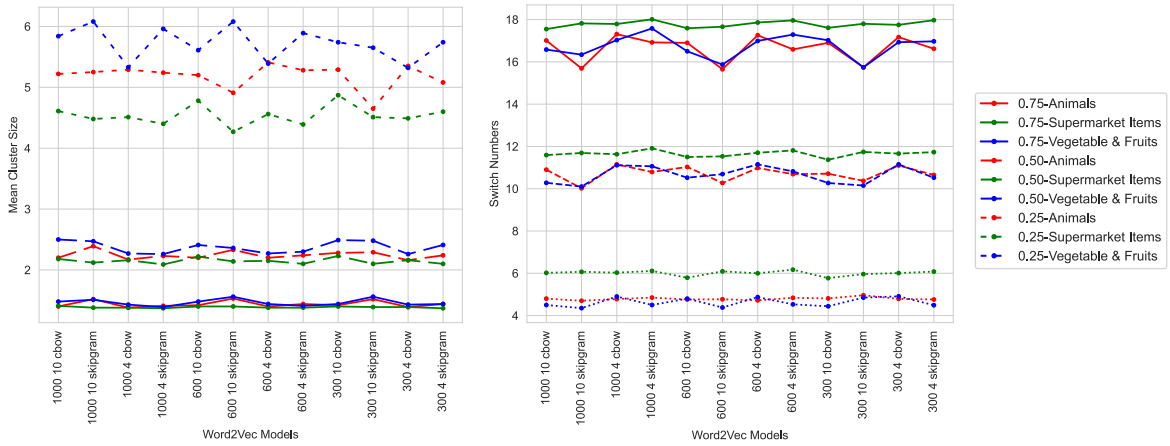


Figure 7.4: Variation in mean cluster size and number of switches according to hyper-parameter selection in Word2vec models

predicted reasonably well, with an F1 score of 0.66 (see Table 7.7), which is just 0.07 points lower than the best vector space model. Thus, the bigram method generates a similar number of switches as manual annotation, but they are placed in slightly different locations. This sufficiently distorts the cluster size to minimise the correlation between predicted and observed cluster sizes.

7.3.4 Group Differences for Manually Annotated versus Automatically Annotated Data

In Tables 7.8–7.10, we report differences between gender and language status groups in the number of switches and mean cluster size. Only the model 600_10_skipgram, not the bigram method or the model 1000_10_skipgram, replicated the significant difference in mean cluster size between genders that we found in the manually annotated data (Animal category) ($p < 0.05$; Table 7.8). No significant differences between categories were observed for Supermarket Items and Fruits & Vegetables.

		N	Mean	Mdn	Max	Min	Std	t-value	p-value	η^2
Number of switches										
Word2vec 600_10_skipgram(0.75)	Gender(M/F)	51/53	15.88/15.57	15.0/15.0	33/46	5/6	5.2/6.53	0.27	0.788	0.053
	Language(Mono/Bi)	60/45	15.48/15.87	15.0/15.0	46/33	5/8	6.44/5.2	-0.324	0.746	-0.064
Word2vec 1000_10_skipgram(0.75)	Gender(M/F)	51/53	16.04/15.49	16.0/15.0	33/47	5/6	5.16/6.71	0.462	0.645	0.091
	Language(Mono/Bi)	60/45	15.72/15.64	15.0/15.0	47/33	5/8	6.61/5.15	0.06	0.952	0.012
Bigram	Gender(M/F)	51/53	11.94/11.25	11.0/11.0	27/36	4/1	5.39/6.15	0.607	0.545	0.119
	Language(Mono/Bi)	60/45	11.17/12.02	10.5/11.0	36/27	1/4	5.97/5.52	-0.743	0.459	-0.147
Mean Cluster size										
Word2vec 600_10_skipgram(0.75)	Gender(M/F)	51/53	1.58/1.48	1.55/1.44	2.33/2.08	1.12/1.06	0.23/0.26	2.023	0.046*	0.397
	Language(Mono/Bi)	60/45	1.51/1.55	1.47/1.54	2.33/2.08	1.06/1.12	0.26/0.23	-0.726	0.469	-0.143
Word2vec 1000_10_skipgram(0.75)	Gender(M/F)	51/53	1.56/1.49	1.54/1.47	2.33/2.25	1.17/1.04	0.22/0.27	1.445	0.152	0.283
	Language(Mono/Bi)	60/45	1.49/1.57	1.48/1.55	2.33/2.25	1.04/1.17	0.24/0.25	-1.64	0.104	-0.323
Bigram	Gender(M/F)	51/53	2.21/2.19	2.0/2.07	5.2/5.5	1.22/1.17	0.72/0.7	0.109	0.914	0.021
	Language(Mono/Bi)	60/45	2.22/2.16	2.08/1.93	5.5/3.43	1.22/1.17	0.78/0.6	0.44	0.661	0.087

Table 7.8: This table presents gender- and language status-based comparisons for Animal category analysed with the bigram method and the two Word2vec models that yielded the best correlation with the Troyer method results, (1) 600-10-skipgram and (2) 1000-10-skipgram, both with 3rd quartile threshold and Snowball stemmer. *$p < .05$

		N	Mean	Mdn	Max	Min	Std	t-value	p-value	η^2
Number of switches										
Word2vec 600_10_skipgram(0.75)	Gender(M/F)	51/53	15.96/15.85	16.0/16.0	32/26	1/5	5.22/4.72	0.113	0.91	0.022
	Language(Mono/Bi)	60/45	15.85/15.89	16.0/16.0	26/32	1/7	4.95/4.98	-0.039	0.969	-0.008
Word2vec 1000_10_skipgram(0.75)	Gender(M/F)	51/53	16.39/16.4	17.0/17.0	32/26	1/6	5.31/4.7	-0.004	0.997	-0.001
	Language(Mono/Bi)	60/45	16.2/16.53	17.0/17.0	26/32	1/7	5.08/4.91	-0.334	0.739	-0.066
Bigram	Gender(M/F)	51/53	9.61/9.38	9.0/9.0	36/16	2/2	4.99/3.19	0.279	0.781	0.055
	Language(Mono/Bi)	60/45	9.37/9.53	10.0/9.0	17/36	2/4	3.49/4.97	-0.2	0.842	-0.039
Mean Cluster size										
Word2vec 600_10_skipgram(0.75)	Gender(M/F)	51/53	1.54/1.59	1.47/1.5	3.0/2.67	1.14/1.11	0.33/0.35	-0.782	0.436	-0.153
	Language(Mono/Bi)	60/45	1.55/1.58	1.43/1.53	3.0/2.56	1.08/1.14	0.36/0.33	-0.406	0.685	-0.08
Word2vec 1000_10_skipgram(0.75)	Gender(M/F)	51/53	1.51/1.52	1.44/1.43	3.0/2.3	1.11/1.11	0.33/0.26	-0.278	0.782	-0.055
	Language(Mono/Bi)	60/45	1.51/1.51	1.42/1.47	3.0/2.3	1.11/1.14	0.32/0.28	0.05	0.96	0.01
Bigram	Gender(M/F)	51/53	2.58/2.63	2.42/2.5	5.0/5.33	1.19/1.69	0.71/0.68	-0.391	0.697	-0.077
	Language(Mono/Bi)	60/45	2.55/2.68	2.42/2.67	5.33/5.0	1.67/1.19	0.65/0.74	-0.901	0.37	-0.178

Table 7.9: This table presents gender- and language status-based comparisons of Fruits & Vegetables category analysed by the bigram model and the two Word2vec models that were selected as the best models for the Animal category data compared to the Troyer method, (1) 600-10-skipgram and (2) 1000-10-skipgram, both with a 3rd quartile threshold and Snowball stemmer.

		N	Mean	Mdn	Max	Min	Std	t-value	p-value	η^2
Number of switches										
Word2vec 600_10_skipgram(0.75)	Gender(M/F)	50/50	18.28/17.14	17.0/18.0	36/28	8/6	5.85/4.93	1.043	0.299	0.209
	Language(Mono/Bi)	58/43	18.05/17.14	17.5/17.0	29/36	6/7	5.12/5.79	0.829	0.409	0.167
Word2vec 1000_10_skipgram(0.75)	Gender(M/F)	50/50	18.4/17.36	17.0/18.0	35/28	8/8	5.78/4.84	0.966	0.337	0.193
	Language(Mono/Bi)	58/43	18.21/17.3	18.0/17.0	30/35	8/8	5.23/5.49	0.833	0.407	0.168
Bigram	Gender(M/F)	50/50	16.76/14.92	15.0/15.0	29/30	9/4	5.46/5.49	1.663	0.099	0.333
	Language(Mono/Bi)	58/43	15.78/15.86	15.0/15.0	27/30	4/5	4.91/6.28	-0.075	0.94	-0.015
Mean Cluster size										
Word2vec 600_10_skipgram(0.75)	Gender(M/F)	50/50	1.43/1.38	1.37/1.31	2.33/2.17	1.0/1.0	0.26/0.24	0.974	0.332	0.195
	Language(Mono/Bi)	58/43	1.37/1.44	1.32/1.42	2.17/2.33	1.07/1.0	0.2/0.3	-1.493	0.139	-0.3
Word2vec 1000_10_skipgram(0.75)	Gender(M/F)	50/50	1.41/1.35	1.36/1.3	2.33/2.0	1.0/1.0	0.23/0.23	1.242	0.217	0.248
	Language(Mono/Bi)	58/43	1.36/1.42	1.33/1.35	1.86/2.33	1.07/1.0	0.18/0.28	-1.381	0.17	-0.278
Bigram	Gender(M/F)	50/50	1.56/1.63	1.54/1.56	2.18/2.83	1.06/1.06	0.29/0.34	-1.097	0.275	-0.219
	Language(Mono/Bi)	58/43	1.58/1.6	1.57/1.53	2.42/2.83	1.06/1.06	0.28/0.37	-0.241	0.81	-0.048

Table 7.10: This table presents gender- and language status-based comparisons for the Supermarket Items category. It includes the bigram model and two Word2vec models, (1) 600-10-skipgram and (2) 1000-10-skipgram, with 3rd quartile threshold and Snowball stemmer. The two Word2vec models were identified as the best models based on comparison to the Troyer method for Animal category data.

7.4 Discussion

With this study, we aimed to create a baseline for Turkish using clustering and switching strategies. The Troyer method is labour-intensive as it necessitates adapting an animal taxonomy into other languages and organising new groups of encountered animal names and requires expertise in the field. Our study is a substantial addition to the existing literature on clustering and switching on Turkish, which has mainly focused on comparing people with neurodegenerative diseases to healthy controls (Çabuk et al., 2020; Kalafatoglu, 2015; Karaca, 2015; Uzgan et al., 2021; Çabuk, 2018). None of these studies shared the adaptation strategy of translating the English animal taxonomy into Turkish, and they do not provide information about how they dealt with local or new animal names that were not in the original taxonomy. This also means that basic normative data on the effects of demographics such as age, gender, language status, or education level on clustering and switching behaviours are unavailable. Our dataset addresses this gap for gender and language status, but unfortunately not for education level or age.

We found no significant differences between genders in total word count, number of perseverations, or number of switches, but there was a small, significant difference in mean cluster size. The findings on gender differences in the literature are equivocal. Several Turkish studies have reported differences in word count between males and females, with males producing more words for animals (Şentürk, 2019), human names (Aki et al., 2022), and meals (Özdemir, 2015)). On the contrary, Jebahi et al. (2022) (Lebanese) and Soriano et al. (2015) (Spanish) found that women produced more items. However, Tuncer (2012), Aki et al. (2022), and İlkmen and Büyükişcan (2022) failed to find a word count difference. This is in line with our results, and with those of Tombaugh et al. (1999) (animals; English), McCarrey et al. (2016) (animals and fruits & vegetables; English), and Gawda and Szepietowska (2013a) (animal and fruits; Polish). Previous works using clustering and switching have either found no difference between gender groups for the animal category (Weiss et al. (2006) for German, Brucki and Rocha (2004) for Italian, and Sokołowski et al. (2020) for Polish) or reported differences in the number of switches (Kosmidis et al. (2004b) (Greek, fruits) and Rosselli et al. (2009) (Spanish, vegetables)). This suggests that the difference in mean cluster size we observed might be a false positive.

As discussed in Chapter 2, there is no bilingual advantage in SVF tasks (Giovannoli et al., 2023) and bilinguals performed similarly well as monolinguals. None of the

most frequent second languages reported in Section 6.9.2 (English, Arabic, Kurdish, and German), were from the same language family as Turkish. Moreover, participants reported proficiency levels between 3.5 and 4 out of 5, indicating that they do not possess near-native competence in their reported second languages. Therefore, the risk of interference between the second language and Turkish was low, which may explain why bilinguals did not perform worse than monolinguals.

While the bigram method performed well for switch location and number of switches, the mean cluster sizes produced by the model did not match the results of manual annotation well. The best performing vector space model, 600_10_skipgram with 0.75 threshold, had superior results across three metrics—switch location, mean cluster size, and number of switches. The 1000_10_skipgram model with 0.75 threshold, which was the best model in our Colombian Spanish study in Chapter 5, also performed well. However, only the best vector space model, 600_10_skipgram, confirmed the existence of a significant difference in cluster size between males and females. For all other comparisons, 600_10_skipgram and 1000_10_skipgram showed no significant differences between genders or between monolinguals and bi-/multilinguals. While the optimal model parameters proved to be relatively similar between Spanish and Turkish, the best performing parameters for Korean and English, as established by (Kim et al., 2019), were different: 300_4_skipgram in English and 1000_4_cbow and 600_10_cbow in Korean. The second best model in Korean and the best model of this study in Turkish are equivalent in dimension (600) and window-size (10) but not architecture. Skipgram is slower than CBOW and is better at finding weak relationships between words in smaller datasets.

7.5 Limitations

One of the most significant limitations of the dataset we gathered as part of the study was the skewed distribution of participants in terms of age and education. This means that we were unable to investigate differences in two demographic factors where differences in performance have been repeatedly reported in the Turkish literature: age and education (socioeconomic status).

Age is a well-researched demographic variable to compare SVF performances. Studies have shown that the number of words produced decreases with increasing age. (Benito-Cuadrado et al., 2002b; Zarino et al., 2014; Llewellyn et al., 2009). Older participants make fewer switches, and their cluster size tends to be larger (Lanting

et al., 2009; Troyer et al., 1997). However, the effect of age itself on SVF performance is comparatively small (Troyer, 2000). The effect of age observed in the Turkish normative studies we found was mixed. While Tuncer (2012) observed a negative effect of age, consistent with the literature, Özdemir (2015) only found an age effect for beverages and famous people, but not for breakfast items, food, or household items. Similarly, Aki et al. (2022) only observed a significant difference in performance for the first name category, not for animals. Şentürk (2019) also failed to find any significant differences between age groups. The vast majority of the participants included in our dataset were between the ages of 18 and 49, therefore, we do not report age comparisons in this study.

Although there were seven levels of education included in our dataset, around 84% of the participants were graduates or postgraduates. This means that there were too few participants with lower education levels to make a meaningful comparison. In the Turkish literature, normative studies agree that education level has a positive effect on SVF performance, regardless of category (Tuncer, 2012; Özdemir, 2015; Şentürk, 2019; Özdemir and Tunçer, 2021; Aki et al., 2022; İlkmen and Büyükişcan, 2022; Tumaç, 1997). Tombaugh et al. (1999) in English and Benito-Cuadrado et al. (2002b) in Spanish provided evidence that participants with higher levels of education perform better in the Animal category, and Nogueira et al. (2016) confirmed this finding for nine different semantic categories in Portuguese. While Da Silva et al. (2004) (Portuguese) and Brincker et al. (2021) (Spanish) found no effect of education on Supermarket products and Fruits, respectively, education was found to have an influence on the Animal category. Furthermore, Da Silva et al. (2004) observed significant group differences in the total number of switches but not the mean cluster size for sequences in the Animal category.

As for bilingualism, the proficiency participants reported in their second language as well as the mean age of acquisition of the second language suggest that most bilingual and multilingual participants had Turkish as their dominant language. Moreover, the second language was almost always English. Our findings need to be replicated for second-generation speakers of Turkish in the Turkish diaspora and more balanced recruitment across the diaspora would be beneficial.

While the average word count across all categories was 25, it is noteworthy that the highest word count of 51 was reached in the Animal category (Fruits & Veg, Max = 44; Supermarket Items, Max = 46). There are many potential reasons for the particularly high number of animals that were named. For example, some individuals might

be professionally involved in animal husbandry or have received an education in veterinary sciences. Alternatively, they could simply be hobbyists with a keen interest in animal species and an accordingly extensive word memory. It is also possible that the maximum word counts reported in Table 7.1 are due to cheating, given that the task was administered remotely. Moreover, it is possible that some participants retested themselves, because there was no explicit countermeasure to prevent this. Data such as location (IP address) and device information (MAC address) were not collected to preserve participant privacy. This could be seen as a disadvantage of online data collection methods, as cheating is not a possibility in traditional methods where data collection is under clinical or experimenter control. In order to improve the detection of such behaviour, future studies could implement an advanced authentication system to create an encrypted timestamp using unique network flow features (IP and MAC addresses) and establishing systems to prevent repeated entries from the same device (e.g. computer, mobile phone, tablet, etc.). However, such solutions are difficult and expensive to implement, whereas our online data collection system was aimed at rapid and cost-effective data gathering. Given that participants in this study did not receive any compensation, their primary motivation for participation was presumably to contribute to science. Consequently, it can be reasonably expected that participants are not likely to take advantage of the opportunity to make multiple submissions. From this perspective, the most effective countermeasure is to explicitly state the purpose of the test in the participant information section, emphasising the importance of participating in the study without prior preparation, acknowledging the inherent challenges in word recall, and advising that participants refrain from retaking the test to ensure the reliability of the results.

7.6 Conclusion

In this chapter, we reported a transparent, replicable analysis of the dataset described in Chapter 6. Computational analysis methods were able to replicate manual annotation results well, with the computationally more expensive vector space model outperforming the simpler bigram model. Due to the limitations of our dataset, we were unable to establish results for key demographic categories such as age and education or socioeconomic status. Since our analyses were all generated using open source software, and the taxonomy used for annotation is provided in this thesis, we have provided a solid foundation for future replications of our work with a more balanced and representative

dataset. The source code for the analysis can be found in a GitHub repository [5].

[5]The code for the semantic verbal fluency analysis `https://github.com/rykostas/Turkish-semantic-verbal-fluency-analysis`

Chapter 8

Conclusion

In this thesis, we aimed to adapt semi-automatic computational linguistics techniques to Turkish SVF sequences in light of previous studies in the SVF field. Consequently, **(1)** a systematic literature review presenting the existing data on Turkish has been produced (Chapter 4), **(2)** a baseline analysis system has been built based on Colombian Spanish SVF data (Chapter 5), **(3)** a fully-online participant centered data collection method has been performed (Chapter 6), and **(4)** a comprehensive linguistic analysis of Turkish SVF sequences has been investigated (Chapter 7).

In this section, we will delve into the outcomes of our research, highlighting our contributions to the field. We will also address our limitations and outline future research direction.

8.1 Findings and Contributions

Chapter 4 addressed the research question (RQ) posed by the systematic review study, in which we aimed to unearth the SVF studies that have been conducted on native Turkish speakers. This review is important in that it presents the normative studies carried out, the commonly used categories, and the scoring techniques used for SVF performance evaluation. Our findings showed that except for two recently published normative studies (Aki et al., 2022; İlkmen and Büyükişcan, 2022), all previous normative SVF studies on Turkish speakers have been published in the Turkish language and have mostly been unpublished theses, which makes them difficult to find for English-speaking researchers. Additionally, no open access normative data has been shared on any platform and clustering and switching methods have not been reported in any normative study. The studies which used clustering and switching based on a

Troyer taxonomy used manually derived features, and no studies have used algorithm-based approaches.

Villalobos et al. (2022) summarised normative data studies in the field of SVF published in various languages, but the study did not present any studies with a Turkish sample. Our study allows researchers to easily find Turkish normative studies and access the details of studies published in Turkish.

It should be noted that SVF is frequently used as an auxiliary tool to monitor diseases in the Turkish literature. Thus, we assumed that the awareness of Turkish norms would increase and that normative data would be easily reached for use in interlingual comparisons (specifically other Turkic languages such as Azerbaijan and Turkmen).

Chapter 5 addressed the RQ posed by our SVF analysis study of Spanish-speaking Colombian Alzheimer's disease patients which we aimed to validate two computational linguistics approaches: vector space models in Section 3.5 and bigram models in Section 3.2 and investigated how they replicate results obtained with the Troyer method in Section 2.4.2.

Our study, which used a dataset from a rarely studied type of AD, familial AD caused by an E280A presenilin-1 mutation, obtained results similar to those of many studies in the field that have focused on patients with non-familial sporadic AD (Raoux et al., 2008; Bertola et al., 2014; Haugrud et al., 2011).

We compared the prediction of switch location by the computational models and the Troyer method. The bigram method achieved $F1{=}0.75$ and Word2vec exceeded $F1{=}0.80$ in some models with third quartile threshold. Although the scores are high, the cluster switching outputs of the automatic methods differ from Troyer's, but this variation is not sufficient to alter the overall results.

The Troyer method easily distinguished between healthy and AD patients based on switch numbers but not mean cluster size. The bigram method and best performing vector space model results are in line with those obtained using the Troyer method. The best performing vector space model also found a small significant difference in mean cluster size between healthy controls and patients with familial AD ($p < 0.05$). Our findings show that both computational approaches replicate the Troyer model sufficiently well to enable differentiation of groups.

Chapter 6 addressed the RQ of the Turkish SVF data collection study where we aimed to create an open access dataset which was collected using a we aimed to design

self paced fully online application. As an alternative to traditional paper-and-pencil or telephone-based data collection methods, in our study, we designed a data collection tool that allows participants to progress at their own pace, and do not to need a practitioner. We presented our designed application in a step-by-step with giving a detailed structure. Using this application, we gathered up-to-date SVF data from individuals living both within and outside of Türkiye whose native language is Turkish. We shared participant's detailed demographic information as well as language ability characteristics. To the best of our knowledge, our study is the first entirely online SVF data collection tool for the Turkish language. Internationally, we found only one online data collection study has been conducted by Cho et al. (2021) for letter fluency, focusing primarily on data analysis and not providing a reusable application with open-source code. In our study, we share the code for the application we designed, making it publicly available and replicable by other researchers. The source code can be found in a github repository [1]. The categories and features (duration, instructions etc.) that we used in our study can be changed by other researchers and application can be adapted into other languages easily. The dataset itself (transcriptions and annotations only) will also be shared with other researchers on request.

Although our study shares similar limitations with other online data collection tools (such as technology requirements, bias towards a younger and wealthier, etc.), also offers benefits, including the opportunity for participants to collect data independently, regardless of time and location. One of the most significant features we present in our study is the extended data confidentiality by storing participants' demographic data on Qualtrics and SVF voice recordings on Google Cloud systems separately.

Chapter 7 addressed the RQ of the Turkish SVF analysis study, in which we aimed to investigate the computational methods we had validated for Spanish in Chapter 5 and determine whether they were successful at replicating the Troyer method for Turkish SVF data. We evaluated the performance of these computational methods by comparing with the Troyer method by gender and bilingualism.

In this study, we collected SVF data for the three most used categories in the Turkish literature: Animal, Fruits & Vegetables, and Supermarket Items. We adapted the Animal category into the Troyer method and compared the results those from with algorithm-based methods to see how the latter replicate internal structure. Our results showed that the vector space model and bigram method are successful in replicat-

[1]The codes of audio recording application `https://github.com/rykostas/SVF`

ing the manually annotated internal structure of clusters and switches. Although both methods are good at finding the locations of switches, the best performing vector space model outperformed the bigram method, with a maximum F1 of 0.73 compared to the bigram's 0.66. Both methods yielded lower scores than a validation study. The best performing models were different in the two studies; the second best model in the Turkish analysis study was the best model for the Colombian Spanish study.

Based on the two best performing vector space models, we presented an in-depth analysis of SVF data for three categories in terms of gender and language status. In the gender comparison, almost no models found any significant difference between genders in any feature or category. The only significant difference between genders was found in the Animal category by the model 600_10_skipgram. It was based on mean cluster size, confirming the finding from the manual analysis that men create larger clusters. Also, in line with results based on manual annotation, there were no differences between monolingual and bi- or multilingual speakers. Therefore, even though the match between computational methods and manual annotation was lower than for the Spanish study, the computational analyses yielded similar results to the manual analyses for Turkish data.

We have shared the code for the analysis of Turkish SVF sequences in this study, making it publicly available and replicable by other researchers. The source code can be found in a GitHub repository [2].

8.2 Limitations

The limitations of the studies are presented in detail in the relevant chapters, but there are some general boundaries discussed below which could affect the quality of a study or reduce its scope.

Dataset: The success of the method employed will vary depending on the scope and representativeness of the collected data. It is important to note that each dataset yields unique results. Extending the results of our Turkish analysis study with normative data which represents the population better would contribute to establishing a more robust baseline for computational linguistics tools for the Turkish language.

Online data collection methods can facilitate reaching larger audiences. However,

[2]The codes of semantic verbal fluency analysis `https://github.com/rykostas/Turkish-semantic-verbal-fluency-analysis`

the scope of our Turkish dataset was narrow in terms of age and level of education, and the foreign data is mostly from the UK. Offering compensation for participation and utilising participant recruiting applications such as Prolific Academic are both viable ways to increase participation, especially by participants residing outside of Türkiye. To include elderly participants, studies can be promoted in places such as retirement social groups or clubs, and assistance can be sought from professional organisations. Particularly, active Facebook groups should be explored and used for promoting the studies.

Computational linguistics tools and requirements: In our study, Turkish Wikipedia articles were used to create a vector space model for Word2vec. Using more recent versions of these articles or expanding the corpus with different text collections could contribute to the creation of models that are more successful in establishing word relationships. The dependence of the model success on the corpus points to another limitation of our study.

In order to handle inflectional words using normalisation libraries (lemmatisation and stemming) are other factors that can impact the results. Compared to more widely spoken languages like Spanish or English, there are few natural language processing tools available for Turkish, which forces a selection from a limited number of available tools. In the future, the experiments in this study should be repeated using multilingual normalisation tools that include the Turkish language or more comprehensive Turkish-specific tools which have yet to be developed.

Baseline method: Our study was based on Troyer's clustering-switching strategy built around an animal taxonomy. The Troyer method is widely accepted as a reference point in the field, and we consider to be the gold standard for computational linguistics methods. However, one of the major limitations of this method is that it was developed for the English language, requiring adaptation and translation for use in different languages. Even for an English dataset, researchers may need to manually place an animal name that was not originally included in the taxonomy. Furthermore, participants may mention extinct animals, like dinosaurs, dodos, and mammoths, or mythical creatures, like unicorns and dragons. Decisions such as creating new groups for these animals or completely eliminating them need to be made by researchers. Direct translation into different languages is also not a definitive solution. For example, in Troyer's original taxonomy, 'pig' is listed as a farm animal, and automatic translation

tools translate it as ‘domuz’ (pig) in Turkish. However, it might be more appropriate to classify it as a wild animal, since mostly Turkish people use ‘domuz’ to refer to ‘boar’. Nevertheless, the ambiguity surrounding whether the person is referring to a farm animal or a wild animal will still persist. Therefore, some manual explanations and annotations are still needed for the Troyer method, and this appears to be a limitation that affects the results. For this purpose, careful documentation of the taxonomies created and the rules followed in future studies will be useful for other researchers.

8.3 Future Directions

In this section, we will outline potential directions for future research addressing two aspects: (1) improving the dataset, and (2) improving the automatic analysis of SVF data.

8.3.1 Future Studies to Address the Gaps in the Existing Work

We will propose two future research directions aimed at addressing the shortcomings of the study we were conducted. First, understanding users’ privacy concerns would reduce the high abandonment rates during the data collection process. The second suggestion is directed at improving the data transcription process we employed for the data analysis and alleviating the burden of manual annotation.

User experience research can help to decrease the number of drop outs during data collection. One of the major challenges in online data collection is the tendency for participants to abandon the study at various points during the data collection process. In our study, we observed that particularly large proportion of individuals left the system after providing demographic information but before beginning the audio recording phase (Chapter 6). Future studies need to investigate the reasons for participants dropping out, which may range from technical issues to potential reservations about providing voice data online. In this regard, it could be beneficial to determine what specific privacy concerns participants have when engaging in an online study. In-depth user studies and thorough tests should be conducted on different hardware configurations and in both mobile and desktop environments; this was, unfortunately, beyond the scope of this thesis.

Improved post-processing systems can streamline the manual transcription control phase. The audio recordings obtained from the data collection study (Chapter 6) were transcribed with Google Speech-to-Text API. However, empty recordings can be detected before passing data to the API, and it might also be possible to automatically detect and remove helpers' speech and instances of participants thinking aloud. Factors that affect audio quality, such as the quality of the recording device, the acoustic environment, and the participant's dialect and accent, can be eliminated with noise reduction methods and the recording can be transcribed better. Moreover, there are limited resources for the Turkish language, which leads to higher word error rates when data is analysed with the current systems compared to analyses of data for languages with a large number of speakers. In order to increase the quality of transcribed data, better automatic speech recognition systems which were developed specifically for the Turkish language should be explored.

8.3.2 Improving Automatic Analysis of SVF Data

In this thesis, the aim was to adapt commonly used computational methods, specifically Word2Vec, to Turkish and to establish a baseline for these methods, which is lacking in the Turkish literature. However, our experiments have been conducted based on lexical features. In addition to the methods that we used, different methods that have been researched in the SVF field were outlined in Chapters 2 and 3. In this context, we propose two future research directions that will enhance the broader applicability of SVF tests by providing in-depth and comprehensive performance evaluations. The first is aimed at implementing the dynamic word embedding methods for Turkish dataset. The second proposed direction concerns evaluating SVF performance with a broader range of features by encompassing not only lexical but also speech-based features.

Dynamic contextualised word embeddings can capture word relations based on a given context and increase the ability to differentiate compared groups. Chapters 5 and 7 provide a detailed examination of the Word2vec method, which is a computational linguistic tool that is widely used in the field. However, Word2vec creates static word representations that can only be changed by retraining the model on a new corpus. Apart from SVF, almost all recent work in computational linguistics has used dynamic contextualised word embeddings that can be adapted for a specific task using the fine-tuning method and can have different representations depending on the context

in which a word is used. An important advantage of using dynamic word embedding methods is the use of existing pre-trained models, which saves the researcher the high costs of training large models.

In the Troyer method, the word 'dog' belongs to both the 'canines' and 'pets' animal categories. When clusters are created, 'dog' is placed in the group where it forms the larger cluster. In Word2vec, there is a vector output for 'dog' and its meaning; whether it is part of 'canines' or 'pets' is determined by one unique vector value. However, in dynamic word embedding methods, 'dog' can have both vector values based on the words used before and after it, and the most suitable vector output is determined based on the relationship between words. Even if it is unclear which retrieval strategy was used by the participant for this word, this method can provide a more effective understanding of where a person made a switch.

In future work, the different context-based methods described in Section 3.4.3 (BERT, RoBERTa, DistilBERT, etc.) should be applied to the data we collected in this study. The early and limited examples in the field have shown that context-based methods have promising results. In their study on a German SVF dataset, Alaçam et al. (2022) compared various static and context-based word embedding methods in different semantic categories. The ConceptNet (Liu and Singh, 2004) contextual method achieved high scores in capturing manually extracted clusters. A study by Nighojkar et al. (2022) on an English SVF dataset compared models' ability to predict which word a person would say after a target word and emphasised that the RoBERTa model performed better than the static methods; Word2vec and Glove. Turkish is under-resourced language for contextual based SVF analysis and as far as we know no studies have tested large language models.

Speech-based features can help to improve models through in-depth analyses going beyond lexical features. While SVF performance is commonly evaluated by examining the total word count and word relationships, it is also possible to uncover addition information beyond words. For instance, Chen et al. (2020) investigated automatic count-based and cluster-based features as well as three different time-related features: waiting time between switches, cluster duration (start–end), and the time it takes for the first two words to appear within a semantic cluster. Using a support vector model-based classifier, they successfully distinguished healthy elderly individuals from MCI patients. They demonstrated that combining lexical-based features with time-based features could lead to better differentiation between groups. In this regard,

their findings strengthen the idea suggested by Mayr (2002) that assessments of SVF that exclude time features may be incomplete.

Pausing between words is one of the common non-lexical features. However, prosodic features such as pauses, stress, intonation, can also provide insights into a person's cognitive state and contribute to a more comprehensive assessment when combined with other features. For example, anger and other emotional changes are observed in Alzheimer's disease patients and can be audible in a patient's intonation while they list words in an SVF test. Furthermore, speech-based features can be supportive, especially in cases such as early-onset AD or MCI, where word-based features may not be enough to differentiate between deficits and healthy aging.

Bibliography

Aartsen, M. J., Smits, C. H., Van Tilburg, T., Knipscheer, K. C., and Deeg, D. J. (2002). Activity in older adults: cause or consequence of cognitive functioning? a longitudinal study on everyday activities and cognitive performance in older adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 57(2):P153–P162.

Abrahams, S., Leigh, P., Harvey, A., Vythelingum, G., Grise, D., and Goldstein, L. (2000). Verbal fluency and executive dysfunction in amyotrophic lateral sclerosis (als). *Neuropsychologia*, 38(6):734–747.

Abwender, D. A., Swan, J. G., Bowerman, J. T., and Connolly, S. W. (2001). Qualitative analysis of verbal fluency output: Review and comparison of several scoring methods. *Assessment*, 8(3):323–338.

Aichberger, M., Busch, M. A., Reischies, F., Ströhle, A., Heinz, A., and Rapp, M. (2010). Effect of physical inactivity on cognitive performance after 2.5 years of follow-up. *GeroPsych*.

Aiello, E. N., Preti, A. N., Pucci, V., Diana, L., Corvaglia, A., Barattieri di San Pietro, C., Difonzo, T., Zago, S., Appollonio, I., Mondini, S., et al. (2022). The italian telephone-based verbal fluency battery (t-vfb): standardization and preliminary clinical usability evidence. *Frontiers in Psychology*, 13:963164.

Ak, M. A. (2023). Comparative analysis of turkey and russia's public diplomacy on the balkans (example of the russian world foundation and yunus emre institute). *Marmara Üniversitesi Siyasal Bilimler Dergisi*, 11(1):1–22.

Akdemir, E. M. (2021). *Assessment of cognitive functions in patients with chronic pain*. PhD thesis, Eskisehir Osmangazi University.

Akgün, B. (2010). *Examination of alcohol dependents' neuropsychological test performance in the context of treatment motivaton*. PhD thesis, Dokuz Eylül University.

Aki, Ö. E., Alkan, B., Demirsöz, T., Velibaşoğlu, B., Taşdemir, T., Erbaş, S. P., Selvi, K., Ergenç, İ., Barişkin, E., and Özdemir, P. (2022). Effects of Age, Gender and Education on Phonemic and Semantic Verbal. *Türk Psikiyatri Dergisi*, 33(1):53–64.

Akın, A. A. and Akın, M. D. (2007). Zemberek, an open source nlp framework for turkic languages. *Structure*, 10(2007):1–5.

Alaçam, Ö., Schüz, S., Wegrzyn, M., Kißler, J., and Zarrieß, S. (2022). Exploring semantic spaces for detecting clustering and switching in verbal fluency. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 178–191.

Albayrak, E. (2015). *The investigation of people who have or have not general procrastination in terms of executive functioning*. PhD thesis, Halic University.

Alkadhi, K. A. (2018). Exercise as a positive modulator of brain function. *Molecular neurobiology*, 55(4):3112–3130.

Allen, H. A., Liddle, P. F., and Frith, C. D. (1993). Negative features, retrieval processes and verbal fluency in schizophrenia. *The British Journal of Psychiatry*, 163(6):769–775.

Altun, M. B. (2022). *Investigation of verbal fluency skills of individuals with relapse remitting multiple sclerosis*. PhD thesis, İstinye University.

Altunkaynak, Y., Usta, Ş., Ertem, D. H., Köksal, A., Dirican, A. C., and Baybaş, S. (2019). Cognitive functioning and silent neurological manifestations in Behçet's disease with ocular involvement. *Noropsikiyatri Arsivi*, 56(3):173–177.

Amunts, J., Camilleri, J. A., Eick, S. B., Heim, S., and Weis, S. (2020). Executive functions predict verbal fluency scores in healthy participants. *Scientific Reports*, pages 1–11.

Amunts, J., Camilleri, J. A., Eickhoff, S. B., Patil, K. R., Heim, S., von Polier, G. G., and Weis, S. (2021). Comprehensive verbal fluency features predict executive function performance. *Scientific reports*, 11(1):6929.

Anderson, J. A., Mak, L., Keyvani Chahi, A., and Bialystok, E. (2018). The language and social background questionnaire: Assessing degree of bilingualism in a diverse population. *Behavior research methods*, 50:250–263.

Arango-Lasprilla, J. C., Cuetos, F., Valencia, C., Uribe, C., and Lopera, F. (2007). Cognitive changes in the preclinical phase of familial alzheimer's disease. *Journal of clinical and experimental neuropsychology*, 29(8):892–900.

Ardila, A. (2020). A cross-linguistic comparison of category verbal fluency test (animals): A systematic review. *Archives of Clinical Neuropsychology*, 35(2):213–225.

Ardila, A., Ostrosky-Solís, F., and Bernal, B. (2006). Cognitive testing toward the future: The example of Semantic Verbal Fluency (ANIMALS). *International Journal of Psychology*, 41(5):324–332.

Arenaza-Urquijo, E. M., Wirth, M., and Chételat, G. (2015). Cognitive reserve and lifestyle: moving towards preclinical alzheimer's disease. *Frontiers in aging neuroscience*, 7:134.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. (2016). A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Ashford, J. (2005). [p-049]: Memtrax computerized memory test, a one-minute dementia screen. *Alzheimer's & Dementia*, 1:S23–S23.

Auriacombe, S., Grossman, M., Carvell, S., Gollomp, S., Stern, M. B., and Hurtig, H. I. (1993). Verbal fluency deficits in parkinson's disease. *Neuropsychology*, 7(2):182.

Aydın, G. (2017). The Effect Of Clustering Technique On Exploration, Planning And Expression Skills In Writing. *Turkish Studies (Elektronik)*, 12(17).

Aydınoğlu, Ü. (2015). *The prevalence of subthreshold autistic symptoms, effect on theory of mind and comorbidities with psychiatric disorders*. PhD thesis, Atatürk University.

Ayers, M. R., Bushnell, J., Gao, S., Unverzagt, F., Gaizo, J. D., Wadley, V. G., Kennedy, R., and Clark, D. G. (2022). Verbal fluency response times predict incident cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 14(1):e12277.

Aytuna, N. and Çapraz, Y. C. (2018). Uses and gratifications of internet use among the elderly in turkey. *Athens Journal of Mass Media and Communications*, 4(2):109–120.

Baker, L. D., Frank, L. L., Foster-Schubert, K., Green, P. S., Wilkinson, C. W., McTiernan, A., Plymate, S. R., Fishel, M. A., Watson, G. S., Cholerton, B. A., et al. (2010). Effects of aerobic exercise on mild cognitive impairment: a controlled trial. *Archives of neurology*, 67(1):71–79.

Balakrishnan, V. and Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performances. *Proceedings of SCEI Seoul Conferences*.

Balcı, L. A. (2016). *Investigation Of Effects Of Cognitive, Balance And Walking Education With Dual Task Training On Fall Risk In Elderly*. PhD thesis, Istanbul Medipol University.

Baldo, J. V., Schwartz, S., Wilkins, D. P., and Dronkers, N. F. (2010). Double dissociation of letter and category fluency following left frontal and temporal lobe lesions. *Aphasiology*, 24(12):1593–1604.

Baldo, J. V. and Shimamura, A. P. (1998). Letter and category fluency in patients with frontal lobe lesions. *Neuropsychology*, 12(2):259.

Baltes, P. B., Staudinger, U. M., Maercker, A., and Smith, J. (1995). People nominated as wise: A comparative study of wisdom-related knowledge. *Psychology and aging*, 10(2):155.

Baran, B. (2008). *Feeling-of-Knowing, Verb Processing and Executive Functions in Non- Demented Parkinson's Disease Patients*. PhD thesis, Bogazici University.

Baysal Kiraç, L. (2012). *Assessment of early cognitive impairment in patients with clinically isolated syndromes and multiple sclerosis*. PhD thesis, Ege University.

Beatty, W. (2002). Fluency in multiple sclerosis: which measure is best? *Multiple Sclerosis Journal*, 8(3):261–264.

Bekris, L. M., Yu, C.-E., Bird, T. D., and Tsuang, D. W. (2010). Genetics of alzheimer disease. *Journal of geriatric psychiatry and neurology*, 23(4):213–227.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Benito-Cuadrado, M., Esteba-Castillo, S., Böhm, P., Cejudo-Bolivar, J., and Peña-Casanova, J. (2002a). Semantic Verbal Fluency of Animals : A Normative and Predictive Study in a Spanish Population. *Journal of Clinical and Experimental Neuropsychology*, 24(8):1117–1122.

Benito-Cuadrado, M., Esteba-Castillo, S., Böhm, P., Cejudo-Bolivar, J., and Peña-Casanova, J. (2002b). Semantic verbal fluency of animals: a normative and predictive study in a spanish population. *Journal of Clinical and Experimental Neuropsychology*, 24(8):1117–1122.

Benton, A. L. (1968). Differential behavioral effects in frontal lobe disease. *Neuropsychologia*, 6(1):53–60.

Berberoğlu, E. (2018). *Investigation of social cognitive functions in familial risk group for psychosis*. PhD thesis, Istanbul Faculty of Medicine.

Bertola, L., Cunha Lima, M. L., Romano-Silva, M. A., de Moraes, E. N., Diniz, B. S., and Malloy-Diniz, L. F. (2014). Impaired generation of new subcategories and switching in a semantic verbal fluency test in older adults with mild cognitive impairment. *Frontiers in Aging Neuroscience*, 6:141.

Beşer, B. (2019). *Evaluation of decision-making capacity in patients with Alzheimer's disease and Dementia with Lewy bodies*. PhD thesis, Istanbul University.

Bettio, L. E., Rajendran, L., and Gil-Mohapel, J. (2017). The effects of aging in the hippocampus and cognitive decline. *Neuroscience & Biobehavioral Reviews*, 79:66–86.

Bialystok, E., Craik, F., and Luk, G. (2008a). "cognitive control and lexical access in younger and older bilinguals": Correction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(4):859–873.

Bialystok, E., Craik, F. I., and Luk, G. (2008b). Lexical access in bilinguals: Effects of vocabulary size and executive control. *Journal of Neurolinguistics*, 21(6):522–538.

Binetti, G., Magni, E., Cappa, S. F., Padovani, A., Bianchetti, A., and Trabucchi, M. (1995). Semantic memory in alzheimer's disease: An analysis of category fluency. *Journal of clinical and experimental neuropsychology*, 17(1):82–89.

Bingöl, A., Eroğlu, G., and Haktanır, I. (1994). Türk toplumunda sözel akıcılık becerisi; bir standardizasyon çalışması. *Ulusal Nöroloji Kongresi*.

Bird, S. and Loper, E. (2004). NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annu. Rev. Psychol.*, 55:803–832.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Booth, L. N. and Brunet, A. (2016). The aging epigenome. *Molecular cell*, 62(5):728–744.

Bora, E., Can, G., Ildız, A., Ulas, G., Ongun, C. H., Inal, N. E., and Ozerdem, A. (2019). Neurocognitive heterogeneity in young offspring of patients with bipolar disorder: The effect of putative clinical stages. *Journal of Affective Disorders*, 257(1606):130–135.

Borkowski, J. G., Benton, A. L., and Spreen, O. (1967). Word fluency and brain damage. *Neuropsychologia*, 5(2):135–140.

Bosnjak, M. and Tuten, T. L. (2001). Classifying response behaviors in web-based surveys. *Journal of Computer-Mediated Communication*, 6(3):JCMC636.

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *The Journal of General Psychology*, 49(2):229–240.

Bousfield, W. A. and Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *The Journal of General Psychology*, 30(2):149–165.

Boyle, R., Jollans, L., Rueda-Delgado, L. M., Rizzo, R., Yener, G. G., McMorrow, J. P., Knight, S. P., Carey, D., Robertson, I. H., Emek-Savas, D. D., Stern, Y., Kenny, R. A., and Whelan, R. (2021). Brain-predicted age difference score is related to specific cognitive functions: a multi-site replication analysis. *Brain imaging and behavior*, 15(1):327–345.

Bozdemir, M. (2008). *To determine the relationship between the Pyramid and Pine Trees Test and the category fluency test*. PhD thesis, Maltepe University.

Brincker, L., Pinheiro, E. M. L., Cera, M. L., and Satler, C. (2021). Semantic verbal fluency analysis in highly educated older adults. *Audiology-Communication Research*, 26.

Broglio, S. P., Eckner, J. T., Paulson, H. L., and Kutcher, J. S. (2012). Cognitive decline and aging: the role of concussive and subconcussive impacts. *Exercise and sport sciences reviews*, 40(3):138.

Brucki, S. M. D. and Rocha, M. S. G. (2004). Category fluency test: effects of age, gender and education on total scores, clustering and switching in brazilian portuguese-speaking subjects. *Brazilian journal of medical and biological research*, 37:1771–1777.

Bunker, L., Hshieh, T. T., Wong, B., Schmitt, E. M., Travison, T., Yee, J., Palihnich, K., Metzger, E., Fong, T. G., and Inouye, S. K. (2017). The sages telephone neuropsychological battery: Correlation with in-person measures. *International journal of geriatric psychiatry*, 32(9):991–999.

Burke, D. M. and Shafto, M. A. (2004). Aging and language production. *Current directions in psychological science*, 13(1):21–24.

Bushnell, J., Unverzagt, F., Wadley, V. G., Kennedy, R., Del Gaizo, J., and Clark, D. G. (2023). Post-processing automatic transcriptions with naïve bayes for verbal fluency scoring. *Social Science Research Network*.

Butković, A. (2018). Sex difference in written verbal fluency task among adolescents. *Logopedija*, 8(2):39–43.

Butters, N., Granholm, E., Salmon, D. P., Grant, I., and Wolfe, J. (1987). Episodic and semantic memory: A comparison of amnesic and demented patients. *Journal of clinical and experimental neuropsychology*, 9(5):479–497.

Butters, N., Wolfe, J., Granholm, E., and Martone, M. (1986). An assessment of verbal recall, recognition and fluency abilities in patients with huntington's disease. *Cortex*, 22(1):11–32.

Çabuk, T. (2018). *Analysis of verbal fluency skills on mild cognitive impairment patients*. PhD thesis, Anadolu University.

Çabuk, T., Torun, S., and Adapınar, D. Ö. (2020). Quantitative and Qualitative Assessment of Verbal Fluency in Amnestic Mild Cognitive Impairment. *Türk Nöroloji Dergisi*, 26(3).

Çağıl İnal, S. (2019). *Effect of music performance on cognitive functions and its relation with BDNF val66met and COMT val158met polymorphisms*. PhD thesis, Ankara University.

Çakar, M. M. (2020). *Evaluation of optical coherance tomography results and cognitive functions in epilepsy patients*. PhD thesis, Trakya University.

Campbell, D. (2021). Normative data. In *Encyclopedia of autism spectrum disorders*, pages 3194–3194. Springer.

Capitani, E., Laiacona, M., and Barbarotto, R. (1999). Gender affects word retrieval of certain categories in semantic fluency tasks. *Cortex*, 35(2):273–278.

Castanho, T. C., Amorim, L., Zihl, J., Palha, J. A., Sousa, N., and Santos, N. C. (2014). Telephone-based screening tools for mild cognitive impairment and dementia in aging studies: a review of validated instruments. *Frontiers in aging neuroscience*, 6:16.

Cegolon, A. and Jenkins, A. (2022). Older adults, cognitively stimulating activities and change in cognitive function. *International Journal of Lifelong Education*, 41(4-5):405–419.

Cevik, N., Köksal, A., Doğan, V. B., Dirican, A. C., Bayramoğlu, S., Özturk, M., and Baybas, S. (2016). Evaluation of cognitive functions of juvenile myoclonic epileptic patients by magnetic resonance spectroscopy and neuropsychiatric cognitive tests concurrently. *Neurological Sciences*, 37(4):623–627.

Chasles, M.-J., Tremblay, A., Escudier, F., Lajeunesse, A., Benoit, S., Langlois, R., Joubert, S., and Rouleau, I. (2020). An examination of semantic impairment in amnestic mci and ad: What can we learn from verbal fluency? *Archives of Clinical Neuropsychology*, 35(1):22–30.

Chen, L., Asgari, M., Gale, R., Wild, K., Dodge, H., and Kaye, J. (2020). Improving the assessment of mild cognitive impairment in advanced age with a novel multi-feature automated speech and language analysis of verbal fluency. *Frontiers in Psychology*, 11:535.

Cho, S., Nevler, N., Parjane, N., Cieri, C., Liberman, M., Grossman, M., and Cousins, K. A. (2021). Automated analysis of digitized letter fluency data. *Frontiers in Psychology*, 12:654214.

Choi, J. and Lee, S.-W. (2020). Improving fasttext with inverse document frequency of subwords. *Pattern Recognition Letters*, 133:165–172.

Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Çilden, E. K. (2006). Stemming turkish words using snowball.

Connick, P., Kolappan, M., Bak, T. H., and Chandran, S. (2012). Verbal fluency as a rapid screening test for cognitive impairment in progressive multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 83(3):346–347.

Correro, A. N. and Nielson, K. A. (2020). A review of minority stress as a risk factor for cognitive decline in lesbian, gay, bisexual, and transgender (lgbt) elders. *Journal of Gay & Lesbian Mental Health*, 24(1):2–19.

Covidence (2013). Systematic review management software.

Cruise, K., Bucks, R., Loftus, A., Newton, R., Pegoraro, R., and Thomas, M. (2011). Exercise and parkinson's: benefits for cognition and quality of life. *Acta Neurologica Scandinavica*, 123(1):13–19.

Çukurova, M. (2020). *The investigation of cognitive functions and clinical high risk status for psychosis in first-degree relatives of patients with substance induced psychotic disorder*. PhD thesis, Sağlık Bilimleri University.

Da Silva, C. G., Petersson, K. M., Faísca, L., Ingvar, M., and Reis, A. (2004). The effects of literacy and education on the quantitative and qualitative aspects of semantic verbal fluency. *Journal of clinical and experimental neuropsychology*, 26(2):266–277.

Dassanayake, T. L., Hewawasam, C., Baminiwatta, A., and Ariyasinghe, D. I. (2021). Regression-based, demographically adjusted norms for victoria stroop test, digit span, and verbal fluency for sri lankan adults. *The Clinical Neuropsychologist*, 35(sup1):S32–S49.

Dayıoğlu, M. and Kırdar, M. G. (2022). Keeping kids in school and out of work: Compulsory schooling and child labor in turkey. *Journal of Human Capital*, 16(4):526–555.

De Smedt, T. and Daelemans, W. (2012). Pattern for python. *The Journal of Machine Learning Research*, 13(1):2063–2067.

Deary, I. J., Corley, J., Gow, A. J., Harris, S. E., Houlihan, L. M., Marioni, R. E., Penke, L., Rafnsson, S. B., and Starr, J. M. (2009). Age-associated cognitive decline. *British medical bulletin*, 92(1):135–152.

Delis, D., Kaplan, F., and Kramer, J. (2001a). *Delis-Kaplan Executive Functions System*. The Psychological Corporation., San Antonio, TX.

Delis, D. C., Kaplan, E., and Kramer, J. H. (2001b). Delis-kaplan executive function system. *Assessment*.

Dementia UK (2020). Types and symptoms of dementia.

Demir, B. and Uluğ, B. (2002). Neuropsychological functions in early and late onset alcoholism. *Türk Psikiyatri Dergisi= Turkish Journal of Psychiatry*, 13(1):15–21.

Demiray, D. Y. and Ertan, S. (2023). Evaluation of cognitive functions in patients with essential tremor, parkinson's disease and combination of essential tremor-parkinson's disease. *Nobel Medicus Journal*, 19(1).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dharma, E. M., Gaol, F. L., Warnars, H., and Soewito, B. (2022). The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification. *J Theor Appl Inf Technol*, 100(2):31.

Di San Pietro, C. B., Luzzatti, C., Ferrari, E., de Girolamo, G., and Marelli, M. (2023). Automated clustering and switching algorithms applied to semantic verbal fluency

data in schizophrenia spectrum disorders. *Language, Cognition and Neuroscience*, pages 1–16.

Di San Pietro, C. B., Marelli, M., and Reverberi, C. (2021). Moving from human ratings to word vectors to classify people with focal dementias: Are we there yet? *CLiC-it*.

Diaz-Asper, C., Chandler, C., Turner, R. S., Reynolds, B., and Elvevåg, B. (2021). Acceptability of collecting speech samples from the elderly via the telephone. *Digital Health*, 7:20552076211002103.

Diker, S. (2014). *Association of cognitive impairment with cortical lesion load detected by double inversion recevery (DIR) in clinically isolated syndromes*. PhD thesis, Hacettepe University.

Diker, S., Has, A. C., Kurne, A., Göçmen, R., Oğuz, K. K., and Karabudak, R. (2016). The association of cognitive impairment with gray matter atrophy and cortical lesion load in clinically isolated syndrome. *Multiple Sclerosis and Related Disorders*, 10:14–21.

Dinç, C. (2019). *Predicting the Disease Trajectory of Mild Cognitive Impairment with the Discrepancy between Semantic and Phonemic Fluency Performance*. PhD thesis, Dokuz Eylül University.

Dritschel, B. H., Williams, J., Baddeley, A. D., and Nimmo-Smith, I. (1992). Autobiographical fluency: A method for the study of personal memory. *Memory & cognition*, 20:133–140.

Durrant, P. (2013). Formulaicity in an agglutinating language: The case of turkish. *Corpus Linguistics and Linguistic Theory*, 9(1):1–38.

Eekelaar, C., Camic, P. M., and Springham, N. (2012). Art galleries, episodic memory and verbal fluency in dementia: An exploratory study. *Psychology of Aesthetics, Creativity, and the Arts*, 6(3):262.

Ekin, E. and Çebi, M. (2021). The Relationship Between Cognitive Reserve, Mood State and Emotion Regulation in Healthy Elderly. *Türkiye Klinikleri Sağlik Bilimleri Dergisi*, 6(3).

Er, Z. C. (2014). *Cardiovascular and neurocognitive changes observed in patients who underwent bypass return early to compare the relationship between intraoperative cerebral oximetry values*. PhD thesis, Marmara University.

Erdogan, Ç. (2016). *Effects of antidepressants with different mechanisms of action on cognitive functions, side effects and social functioning*. PhD thesis, Kırıkkale University.

Erk, K. (2012). Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Linguistics and Language Compass*, 6(10):635–653.

Erol, A., Bayram, S., Kosger, F., and Mete, L. (2012). Executive functions in patients with familial versus sporadic schizophrenia and their parents. *Neuropsychobiology*, 66(2):93–99.

Ersan, F. (2014). *Comparison with in terms neurocognitive functions, impulsivity and theory of mind between alcohol dependence diagnosed adults and healthy volunteers*. PhD thesis, Maltepe University.

Esteves, C. S., Oliveira, C. R., Moret-Tatay, C., Navarro-Pardo, E., Carli, G. A. D., Silva, I. G., Irigaray, T. Q., and Argimon, I. I. d. L. (2015). Phonemic and semantic verbal fluency tasks: normative data for elderly brazilians. *Psicologia: Reflexão e Crítica*, 28:350–355.

Estrada-Plana, V., Montanera, R., Ibarz-Estruga, A., March-Llanes, J., Vita-Barrull, N., Guzmán, N., Ros-Morente, A., Ayesa Arriola, R., and Moya-Higueras, J. (2021). Cognitive training with modern board and card games in healthy older adults: two randomized controlled trials. *International Journal of Geriatric Psychiatry*, 36(6):839–850.

Evlice, A. (2016). Effect of demographic data on neuropsychological tests. *Cukurova Medical Journal*, 41(3):528–532.

Fama, R., Rosenbloom, M. J., Sassoon, S. A., Thompson, M. A., Pfefferbaum, A., and Sullivan, E. V. (2011). Remote semantic memory for public figures in hiv infection, alcoholism, and their comorbidity. *Alcoholism: Clinical and Experimental Research*, 35(2):265–276.

Farzanfar, D., Statucka, M., and Cohn, M. (2018). Automated indices of clustering and switching of semantic verbal fluency in parkinson's disease. *Journal of the International Neuropsychological Society*, 24(10):1047–1056.

Fellbaum, C. (2010). Princeton university: About wordnet.

Feng, L., Chong, M. S., Lim, W. S., and Ng, T. P. (2012). The modified mini-mental state examination test: normative data for singapore chinese older adults and its performance in detecting early cognitive impairment. *Singapore Med J*, 53(7):458–62.

Fichman, H. C., Fernandes, C. S., Nitrini, R., Lourenço, R. A., Paradela, E. M. d. P., Carthery-Goulart, M. T., and Caramelli, P. (2009). Age and educational level effects on the performance of normal elderly on category verbal fluency tasks. *Dementia & neuropsychologia*, 3:49–54.

Fjell, A. M., McEvoy, L., Holland, D., Dale, A. M., Walhovd, K. B., Initiative, A. D. N., et al. (2014). What is normal in normal aging? effects of aging, amyloid and alzheimer's disease on the cerebral cortex and the hippocampus. *Progress in neurobiology*, 117:20–40.

Flatt, J. D., Johnson, J. K., Karpiak, S. E., Seidel, L., Larson, B., and Brennan-Ing, M. (2018). Correlates of subjective cognitive decline in lesbian, gay, bisexual, and transgender older adults. *Journal of Alzheimer's Disease*, 64(1):91–102.

Fonseca, R. P., Marcotte, K., Hubner, L. C., Zimmermann, N., Netto, T. M., Bizzo, B., Döring, T., Landeira-Fernandez, J., Gasparetto, E. L., Joanette, Y., et al. (2021). The impact of age and education on phonemic and semantic verbal fluency: Behavioral and fmri correlates. *BioRxiv*, pages 2021–01.

Fossati, P., Ergis, A.-M., Allilaire, J.-F., et al. (2003). Qualitative analysis of verbal fluency in depression. *Psychiatry research*, 117(1):17–24.

Fürnkranz, J. (1998). A Study Using n -gram Features for Text Categorization. *Austrian Research Institute for Artifical Intelligence*, 3:1–10.

Galaverna, F., Bueno, A. M., Morra, C. A., Roca, M., and Torralva, T. (2016). Analysis of errors in verbal fluency tasks in patients with chronic schizophrenia. *The European Journal of Psychiatry*, 30(4):305–320.

Garcia, A. M. (2023). Linguistic markers of alzheimer's in spanish speakers: Automated metrics for free speech and verbal fluency tasks. *Alzheimer's & Dementia*, 19:e062609.

Gawda, B. and Szepietowska, E. (2013a). Impact of unconscious emotional schemata on verbal fluency–sex differences and neural mechanisms. *NeuroQuantology*, 11(3).

Gawda, B. and Szepietowska, E. M. (2013b). Semantic and affective verbal fluency: sex differences. *Psychological Reports*, 113(1):246–256.

Gell, N. M., Rosenberg, D. E., Demiris, G., LaCroix, A. Z., and Patel, K. V. (2015). Patterns of technology use among older adults with and without disabilities. *The Gerontologist*, 55(3):412–421.

Ghasemian-shirvan, E., Shirazi, S. M., Aminikhoo, M., and Zareaan, M. (2018). Preliminary Normative Data of Persian Phonemic and Semantic Verbal Fluency Test. *Iranian Journal of Psychiatry*, 13(4):288–295.

Gilbert, S. J. and Burgess, P. W. (2008). Executive function. *Current Biology*, 18(3):110–114.

Gillen, G. and Rubio, K. B. (2016). Treatment of cognitive-perceptual deficits: A function-based approach. In *Stroke rehabilitation*, pages 612–646. Elsevier.

Giovannoli, J., Martella, D., and Casagrande, M. (2023). Executive functioning during verbal fluency tasks in bilinguals: A systematic review. *International Journal of Language & Communication Disorders*.

Gocer March, E. and Pattison, P. (2006). Semantic verbal fluency in alzheimer's disease: approaches beyond the traditional scoring system. *Journal of Clinical and Experimental Neuropsychology*, 28(4):549–566.

Gökçe, E. (2020). *The effect of open and closed skill exercises on the development of cognitive skills and the level of peripheral protein signals in athletes*. PhD thesis, Ankara University.

Gökçe, E., Güneş, E., Hayme, S., Aslan, E., Asutay, O., Aşar, B., Çetin, M. N., and Çevik, F. (2021). Effects of Playing Tennis on Cognition: A Pilot Study to Examine Hand Preference Effect. *Ankara Üniversitesi Tip Fakültesi Mecmuasi*, 74(1).

Goldberg, E. (1986). Varieties of perseveration: A comparison of two taxonomies. *Journal of Clinical and Experimental Neuropsychology*.

Gollan, T. H., Montoya, R. I., and Werner, G. A. (2002). Semantic and letter fluency in spanish-english bilinguals. *Neuropsychology*, 16(4):562.

Gomez, R. G. and White, A. (2006). Using verbal fluency to detect very mild dementia of the Alzheimer type. *Archives of Clinical Neuropsychology*, 21:771–775.

Gordon, J. K., Young, M., and Garcia, C. (2018). Why do older adults have difficulty with semantic fluency? *Aging, Neuropsychology, and Cognition*, 25(6):803–828.

Grabbe, J. W. (2011). Sudoku and working memory performance for older adults. *Activities, Adaptation & Aging*, 35(3):241–254.

Gregory, S., Linz, N., König, A., Langel, K., Pullen, H., Luz, S., Harrison, J., and Ritchie, C. W. (2022). Remote data collection speech analysis and prediction of the identification of alzheimer's disease biomarkers in people at risk for alzheimer's disease dementia: the speech on the phone assessment (speak) prospective observational study protocol. *BMJ open*, 12(3):e052250.

Gruenewald, P. J. and Lockhead, G. R. (1980). The free recall of category examples. *Journal of Experimental Psychology: Human Learning and Memory*, 6(3):225.

Gücüyener, D., Özdemir, G., and Uzuner, N. (1998). The place of lizuride in aphasia pharmacotheraopy. *Türk Beyin Damar Hastalıklari Dergisi*, 4(1).

Guerra-Carrillo, B., Katovich, K., and Bunge, S. A. (2017). Does higher education hone cognitive functioning and learning efficacy? findings from a large and diverse sample. *PloS one*, 12(8):e0182276.

Gultekin, G., Batun, G., Yürüyen, M., Yavuzer, H., Akcan, F., and Emül, M. (2017). Cognitive impairment and plasma phoenixin level. *European Neuropsychopharmacology*, 27:S1027–S1028.

Gün, F. and Baskan, G. A. (2014). New education system in turkey (4+ 4+ 4): A critical outlook. *Procedia-Social and Behavioral Sciences*, 131:229–235.

Güneş, E., Üstün, S., Gökçe, E., Akat, F., Armağan, E., Gündoğdu, H. E., Bataş, K. K., Ekicioğlu, N., Akkuş, S. N., and Çil, Y. (2022). Effect of sleep duration on working

memory and verbal fluency functions of medical faculty students. *Ankara Universites Tip Fakultesi Mecmuasi= Journal of Ankara University Faculty of Medicine*, 75(4):479.

Gupta, P. and Jaggi, M. (2021). Obtaining better static word embeddings using contextual embedding models. *arXiv preprint arXiv:2106.04302*.

Gürses, N. (2009). *Correlations between neurocognitive performances and schizotypal features in first degree relatives of schizophrenia patients*. PhD thesis, Hacettepe University.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., et al. (2021). Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Hanagasi, H. A., Gürvit, I. H., Ermutlu, N., Kaptanoğlu, G., Karamursel, S., İdrisoğlu, H. A., Emre, M., and Demiralp, T. (2002). Cognitive impairment in amyotrophic lateral sclerosis: Evidence from neuropsychological investigation and event-related potentials. *Cognitive Brain Research*, 14(2):234–244.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Haugrud, N., Crossley, M., and Vrbancic, M. (2011). Clustering and switching strategies during verbal fluency performance differentiate alzheimer's disease and healthy aging. *Journal of the International Neuropsychological Society*, 17(6):1153–1157.

Haugrud, N., Lanting, S., and Crossley, M. (2010). The effects of age, sex and alzheimer's disease on strategy use during verbal fluency tasks. *Aging, Neuropsychology, and Cognition*, 17(2):220–239.

Hazin, I., Leite, G., Oliveira, R. M., Alencar, J. C., Fichman, H. C., Marques, P. d. N., and De Mello, C. B. (2016). Brazilian normative data on letter and category fluency tasks: Effects of gender, age, and geopolitical region. *Frontiers in Psychology*, 7:684.

Helm-Estabrooks, N., Ramage, A., Bayles, K. A., and Cruz, R. (1998). Perseverative behaviour in fluent and non-fluent aphasic adults. *Aphasiology*, 12(7-8):689–698.

Henry, J. D. and Crawford, J. R. (2004). A meta-analytic review of verbal fluency performance following focal cortical lesions. *Neuropsychology*, 18(2):284.

Henry, J. D. and Crawford, J. R. (2005). A meta-analytic review of verbal fluency deficits in depression. *Journal of clinical and experimental neuropsychology*, 27(1):78–101.

Henry, J. D. and Phillips, L. H. (2006). Covariates of production and perseveration on tests of phonemic, semantic and alternating fluency in normal aging. *Aging, Neuropsychology, and Cognition*, 13(3-4):529–551.

Herlitz, A., Nilsson, L.-G., and Bäckman, L. (1997). Gender differences in episodic memory. *Memory & cognition*, 25(6):801–811.

Herrera-García, J. D., Rego-García, I., Guillén-Martínez, V., Carrasco-García, M., Valderrama-Martín, C., Vílchez-Carrillo, R., López-Alcalde, S., and Carnero-Pardo, C. (2019). Discriminative validity of an abbreviated semantic verbal fluency test. *Dementia & Neuropsychologia*, 13:203–209.

Hoffman, P. (2018). An individual differences approach to semantic cognition: Divergent effects of age on representation, retrieval and selection. *Scientific reports*, 8(1):8145.

Holmlund, T. B., Cheng, J., Foltz, P. W., Cohen, A. S., and Elvevåg, B. (2019). Updating verbal fluency analysis for the 21st century: Applications for psychiatry. *Psychiatry Research*, 273:767–769.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hossain, M. R., Hoque, M. M., Siddique, N., and Sarker, I. H. (2021). Bengali text document categorization based on very deep convolution neural network. *Expert Systems with Applications*, 184:115394.

Howard, D. and Patterson, K. E. (1992). The pyramids and palm trees test. *Bury St EdmundsThames Valley Test Company*.

Huang, G., Wharton, W., Travison, T., Ho, M., Gleason, C., Asthana, S., Bhasin, S., and Basaria, S. (2015). Effects of testosterone administration on cognitive function in hysterectomized women with low testosterone levels: a dose–response randomized trial. *Journal of endocrinological investigation*, 38:455–461.

Hurks, P. P. (2012). Does instruction in semantic clustering and switching enhance verbal fluency in children? *The Clinical Neuropsychologist*, 26(6):1019–1037.

Hurks, P. P., Schrans, D., Meijs, C., Wassenberg, R., Feron, F., and Jolles, J. (2010). Developmental changes in semantic verbal fluency: Analyses of word productivity as a function of time, clustering, and switching. *Child Neuropsychology*, 16(4):366–387.

Hyde, J. S. and Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological bulletin*, 104(1):53.

Iizuka, A., Suzuki, H., Ogawa, S., Kobayashi-Cuya, K. E., Kobayashi, M., Takebayashi, T., and Fujiwara, Y. (2019). Can cognitive leisure activity prevent cognitive decline in older adults? a systematic review of intervention studies. *Geriatrics & gerontology international*, 19(6):469–482.

İlkmen, Y. S. and Büyükişcan, E. S. (2022). Verbal fluency tests: Normative data stratified by age and education in an istanbul sample. *Turkish Journal of Neurology/Turk Noroloji Dergisi*, 28(2).

Iñesta, C., Oltra-Cucarella, J., and Sitges-Maciá, E. (2022). Regression-based normative data for independent and cognitively active spanish older adults: verbal fluency tests and boston naming test. *International Journal of Environmental Research and Public Health*, 19(18):11445.

İpekten, E. (2018). *Cognitive functions of the amatuer football players and its relation with indicators of brain damage, neurotrophic factors and myokines*. PhD thesis, Selçuk University.

Isaacs, B. and Kennie, A. T. (1973). The set test as an aid to the detection of dementia in old people. *The British Journal of Psychiatry*, 123(575):467–470.

Ismatullina, V., Voronin, I., Shelemetieva, A., and Malykh, S. (2014). Cross-cultural study of working memory in adolescents. *Procedia-Social and Behavioral Sciences*, 146:353–357.

Istek, O. and Cicekli, I. (2007). A link grammar for an agglutinative language. *Proceedings of Recent Advances in Natural Language Processing (RANLP 2007)*, pages 285–290.

Jebahi, F., Abou Jaoude, R., and Ellis, C. (2022). Semantic verbal fluency task: The effects of age, educational level, and sex in lebanese-speaking adults. *Applied Neuropsychology: Adult*, 29(5):936–940.

Jha, P. and Parvati, P. (2014). Assessing progress on universal elementary education in india: A note on some key constraints. *Economic and Political Weekly*, pages 44–51.

Jiang, W. (2000). The relationship between culture and language. *ELT journal*, 54(4):328–334.

Jones, H. N. (2009). Prosody in parkinson's disease. *Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders*, 19(3):77–82.

Jun, E., Hsieh, G., and Reinecke, K. (2017). Types of motivation affect study selection, attention, and dropouts in online experiments. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–15.

Kalafatoglu, P. (2015). *Verbal fluency performance in alzheimer's disease*. PhD thesis, Anadolu University.

Kandemir, M. (2006). *Cognitive impairment in infratentorial strokes*. PhD thesis, Bakırköy Ruh Sağlığı ve Sinir Hastalıkları Hastanesi.

Kandemir, M., Örnek, İ., and Kirbaş, D. (2009). Cognitive Impairment in Infratentorial Strokes. *Turk Norol Derg*, 15:166–173.

Kapu, A. (2019). *Psychopathological vulnerability among first degree relatives of schizophrenia patients: The relationship between executive functions, metacognition and emotional regulation with psychopathological symptoms*. PhD thesis, Saglik Bilimleri University.

Karaca, E. (2015). *The verbal fluency performances of people with frontotemporal lobar degeneration*. PhD thesis, Anadolu University.

Karahan, M., Kocabeyoğlu, S. S., Kervan, Ü., Sert, D. E., Erdogan Bakar, E., Aygun, E., Tola, M., Demirkan, B., Mungan, S., Catav, Z., and Pac, M. (2021). More continuous flow, better learning? The effect of aortic valve opening in patients with left ventricular assist device. *The International journal of artificial organs*, 44(5):325–331.

Kaur, J. and Buttar, P. K. (2018). A systematic review on stopword removal algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, 4(4):207–210.

Kavé, G. (2005). Phonemic Fluency , Semantic Fluency , and Difference Scores : Normative Data for Adult Hebrew Speakers. *Journal of Clinical and Experimental Neuropsychology*, 27(6):690–699.

Kavé, G., Heled, E., Vakil, E., and Agranov, E. (2011). Which verbal fluency measure is most useful in demonstrating executive deficits after traumatic brain injury ? *Journal of clinical and experimental neuropsychology*, 3395.

Kawano, N., Umegaki, H., Suzuki, Y., Yamamoto, S., Mogi, N., and Iguchi, A. (2010). Effects of educational background on verbal fluency task performance in older adults with alzheimer's disease and mild cognitive impairment. *International Psychogeriatrics*, 22(6):995–1002.

Kaya, B. K. and Alpözgen, A. Z. (2022). Comparing the Cognitive Functioning Effects of Aerobic and Pilates Exercises for Inactive Young Adults: A Randomized Controlled Trial. *Perceptual and motor skills*, 129(1):134–152.

Keijzer, M. C. and Schmid, M. S. (2016). Individual differences in cognitive control advantages of elderly late dutch-english bilinguals. *Linguistic Approaches to Bilingualism*, 6(1-2):64–85.

Kempler, D., Teng, E. L., Dick, M., Taussig, I. M., and Davis, D. S. (1998). The effects of age, education, and ethnicity on verbal fluency. *Journal of the International Neuropsychological Society*, 4(6):531–538.

Khalil, M. S. (2010a). Preliminary Arabic normative data of neuropsychological tests : The verbal and design fluency. *Journal of Clinical and Experimental Neuropsychology*, 32(9):1028–1035.

Khalil, M. S. (2010b). Preliminary arabic normative data of neuropsychological tests: The verbal and design fluency. *Journal of Clinical and Experimental Neuropsychology*, 32(9):1028–1035.

Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., and Rudzicz, F. (2019). A survey of word embeddings for clinical text. *Journal of Biomedical Informatics*, 100:100057.

Kheloui, S., Brouillard, A., Rossi, M., Marin, M.-F., Mendrek, A., Paquette, D., and Juster, R.-P. (2021). Exploring the sex and gender correlates of cognitive sex differences. *Acta Psychologica*, 221:103452.

Kherwa, P. and Bansal, P. (2019). Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, 7(24).

Kiela, D. and Clark, S. (2014). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 21–30.

Kılavuz Ören, G. (2020). *The effects of 8-week aerobic exercise training in individuals with focal epilepsy*. PhD thesis, Pamukkale University.

Kim, H., Kim, J., Kim, D. Y., and Heo, J. (2011). Differentiating between aphasic and nonaphasic stroke patients using semantic verbal fluency measures with administration time of 30 seconds. *European neurology*, 65(2):113–117.

Kim, N., Kim, J.-h., Wolters, M. K., Macpherson, S. E., and Park, J. C. (2019). Automatic Scoring of Semantic Fluency. *Frontiers in Psychology*, 10(May):1–16.

Kintz, S. and Wright, H. H. (2017). Semantic knowledge use in discourse: Influence of age. *Discourse processes*, 54(8):670–681.

Kiraç, L. B., Ekmekçi, Ö., Yüceyar, N., and Kocaman, A. S. (2014). Assessment of early cognitive impairment in patients with clinically isolated syndromes and multiple sclerosis. *Behavioural Neurology*, 2014.

Knopman, D. S., Knudson, D., Yoes, M. E., and Weiss, D. J. (2000). Development and standardization of a new telephonic cognitive screening test: the minnesota cognitive acuity screen (mcas). *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, 13(4):286–296.

Koc, I., Hancioglu, A., and Cavlin, A. (2008). Demographic differentials and demographic integration of turkish and kurdish populations in turkey. *Population research and policy review*, 27(4):447–457.

Kochhann, R., Holz, M. R., Beber, B. C., Chaves, M. L., and Fonseca, R. P. (2018). Reading and writing habits as a predictor of verbal fluency in elders. *Psychology & Neuroscience*, 11(1):39.

König, A., Linz, N., Tröger, J., Wolters, M., Alexandersson, J., and Robert, P. (2018a). Fully automatic speech-based analysis of the semantic verbal fluency task. *Dementia and geriatric cognitive disorders*, 45(3-4):198–209.

König, R., Seifert, A., and Doh, M. (2018b). Internet use among older europeans: an analysis based on share data. *Universal Access in the Information Society*, 17(3):621–633.

Kornfilt, J. (1990). Turkish and the turkic languages. *The world's major languages*, 2.

Kosmidis, M. H., Vlahou, C. H., Panagiotaki, P., and Kiosseoglou, G. (2004a). The verbal fluency task in the Greek population : Normative data , and clustering and switching strategies. *Journal of the International Neuropsychological Society: JINS*, 10(2):164.

Kosmidis, M. H., Vlahou, C. H., Panagiotaki, P., and Kiosseoglou, G. (2004b). The verbal fluency task in the greek population: Normative data, and clustering and switching strategies. *Journal of the International Neuropsychological Society*, 10(2):164–172.

Laisney, M., Matuszewski, V., Mézenge, F., Belliard, S., de la Sayette, V., Eustache, F., and Desgranges, B. (2009). The underlying mechanisms of verbal fluency deficit in frontotemporal dementia and semantic dementia. *Journal of Neurology*, 256:1083–1094.

Lanting, S., Haugrud, N., and Crossley, M. (2009). The effect of age and sex on clustering and switching during speeded verbal fluency tasks. *Journal of the International Neuropsychological Society*, 15(2):196–204.

Lasprilla, J. C. A., Iglesias, J., and Lopera, F. (2003). Neuropsychological stydy of familial alzheimer's disease caused by mutation e280a in the presenilin 1 gene. *American Journal of Alzheimer's Disease & Other Dementias®*, 18(3):137–146.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics*, 3:211–225.

Levy, R. et al. (1994). Aging-associated cognitive decline. *International Psychogeriatrics*, 6(1):63–68.

Lezak, M. D., Howieson, D. B., Loring, D. W., Fischer, J. S., et al. (2004). *Neuropsychological assessment*. Oxford University Press, USA.

Liang, J., Xiao, Y., Zhang, Y., Hwang, S.-w., and Wang, H. (2017). Graph-based wrong isa relation detection in a large-scale lexical taxonomy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Lindsay, H., Mueller, P., Linz, N., Zeghari, R., Mina, M. M., König, A., and Tröger, J. (2021a). Dissociating semantic and phonemic search strategies in the phonemic verbal fluency task in early dementia. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 32–44.

Lindsay, H., Müller, P., Kröger, I., Tröger, J., Linz, N., König, A., Zeghari, R., Verhey, F. R., and Ramakers, I. H. (2021b). Multilingual learning for mild cognitive impairment screening from a clinical speech task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 830–838.

Linz, N., Tröger, J., Alexandersson, J., and Konig, A. (2017a). Using neural word embeddings in the analysis of the clinical semantic verbal fluency task. In *IWCS 2017-12th International Conference on Computational Semantics*, pages 1–7.

Linz, N., Tröger, J., Alexandersson, J., Wolters, M., König, A., and Robert, P. (2017b). Predicting dementia screening and staging scores from semantic verbal fluency performance. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 719–728. IEEE.

Lison, P. and Kutuzov, A. (2017). Redefining context windows for word embedding models: An experimental study. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 284–288.

Liu, H. and Singh, P. (2004). Conceptnet—a practical commonsense reasoning toolkit. *BT technology journal*, 22(4):211–226.

Llewellyn, D. J., Matthews, F. E., Function, T. M. R. C. C., and CFAS), A. S. M. (2009). Increasing levels of semantic verbal fluency in elderly english adults. *Aging, Neuropsychology, and Cognition*, 16(4):433–445.

Lopera, F., Ardilla, A., Martínez, A., Madrigal, L., Arango-Viana, J. C., Lemere, C. A., Arango-Lasprilla, J. C., Hincapié, L., Arcos-Burgos, M., Ossa, J. E., et al. (1997). Clinical features of early-onset alzheimer disease in a large kindred with an e280a presenilin-1 mutation. *Jama*, 277(10):793–799.

Lopes, M., Brucki, S. M. D., Giampaoli, V., and Mansur, L. L. (2009a). Semantic verbal fluency test in dementia: preliminary retrospective analysis. *Dementia & Neuropsychologia*, 3:315–320.

Lopes, M., Maria, S., Brucki, D., Giampaoli, V., and Mansur, L. L. (2009b). Semantic Verbal Fluency test in dementia Preliminary retrospective analysis. *Dementia & Neuropsychologia*, 3(4):315–320.

López-Higes, R., Rubio-Valdehita, S., Fernández-Blázquez, M. A., Lojo-Seoane, C., Ávila-Villanueva, M., Montenegro-Peña, M., Mallo, S. C., and Delgado-Losada, M. L. (2022). Spanish consortium for ageing normative data (scand): semantic verbal fluency tests. *Archives of Clinical Neuropsychology*, 37(2):352–364.

Lövdén, M., Fratiglioni, L., Glymour, M. M., Lindenberger, U., and Tucker-Drob, E. M. (2020). Education and cognitive functioning across the life span. *Psychological Science in the Public Interest*, 21(1):6–41.

Luk, G. and Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of cognitive Psychology*, 25(5):605–621.

Lundin, N. B., Jones, M. N., Myers, E. J., Breier, A., and Minor, K. S. (2022). Semantic and phonetic similarity of verbal fluency responses in early-stage psychosis. *Psychiatry research*, 309:114404.

Lundin, N. B., Todd, P. M., Jones, M. N., Avery, J. E., O'Donnell, B. F., and Hetrick, W. P. (2020). Semantic search in psychosis: Modeling local exploitation and global exploration. *Schizophrenia bulletin open*, 1(1):sgaa011.

Luo, L., Luk, G., and Bialystok, E. (2010). Effect of language proficiency and executive control on verbal fluency performance in bilinguals. *Cognition*, 114(1):29–41.

MacPherson, S. E., Parra, M. A., Moreno, S., Lopera, F., and Della Sala, S. (2012). Dual task abilities as a possible preclinical marker of alzheimer's disease in carriers

of the e280a presenilin-1 mutation. *Journal of the International Neuropsychological Society*, 18(2):234–241.

MacPherson, S. E., Parra, M. A., Moreno, S., Lopera, F., and Della Sala, S. (2015). Dual memory task impairment in e280a presenilin-1 mutation carriers. *Journal of Alzheimer's Disease*, 44(2):481–492.

Mandolesi, L., Polverino, A., Montuori, S., Foti, F., Ferraioli, G., Sorrentino, P., and Sorrentino, G. (2018). Effects of physical exercise on cognitive functioning and wellbeing: biological and psychological benefits. *Frontiers in psychology*, 9:509.

Marceaux, J. C., Prosje, M. A., McClure, L. A., Kana, B., Crowe, M., Kissela, B., Manly, J., Howard, G., Tam, J. W., Unverzagt, F. W., et al. (2019). Verbal fluency in a national sample: Telephone administration methods. *International journal of geriatric psychiatry*, 34(4):578–587.

Mardani, N., Pourjafari, M., Irandegani, M. A., Ahmadi, N., and Baghban, K. (2020). The effect of bilingualism on the processing of clustering and switching in verbal fluency tasks. *Journal of Rehabilitation Sciences & Research*, 7(3):114–117.

Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (leap-q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*.

Marsh, J. E., Hansson, P., Sörman, D. E., and Ljungberg, J. K. (2019). Executive processes underpin the bilingual advantage on phonemic fluency: evidence from analyses of switching and clustering. *Frontiers in psychology*, 10:1355.

Maruff, P., Thomas, E., Cysique, L., Brew, B., Collie, A., Snyder, P., and Pietrzak, R. H. (2009). Validity of the cogstate brief battery: relationship to standardized tests and sensitivity to cognitive impairment in mild traumatic brain injury, schizophrenia, and aids dementia complex. *Archives of Clinical Neuropsychology*, 24(2):165–178.

Mathuranath, P., George, A., Cherian, P., Alexander, A. l., Sarma, S., and Sarma, P. (2003). Effects of age, education and gender on verbal fluency. *Journal of clinical and experimental neuropsychology*, 25(8):1057–1064.

Mathuranath, P., Nestor, P., Berrios, G., Rakowicz, W., and Hodges, J. (2000a). A brief cognitive test battery to differentiate alzheimer's disease and frontotemporal dementia. *Neurology*, 55(11):1613–1620.

Mathuranath, P. S., Nestor, P. J., Berrios, G. E., Rakowicz, W., and Hodges, J. R. (2000b). A brief cognitive test battery to differentiate Alzheimer's disease and frontotemporal dementia. *Neurology*, pages 315–320.

Mayr, U. (2002). On the dissociation between clustering and switching in verbal fluency: Comment on troyer, moscovitch, winocur, alexander and stuss. *Neuropsychologia*, 40(5):562–566.

Mayr, U. and Kliegl, R. (2000). Complex semantic processing in old age: Does it stay or does it go? *Psychology and aging*, 15(1):29.

McCann, B., Bradbury, J., Xiong, C., and Socher, R. (2017). Learned in translation: Contextualized word vectors. *Advances in neural information processing systems*, 30.

McCarrey, A. C., An, Y., Kitner-Triolo, M. H., Ferrucci, L., and Resnick, S. M. (2016). Sex differences in cognitive trajectories in clinically normal older adults. *Psychology and aging*, 31(2):166.

McDowd, J., Hoffman, L., Rozek, E., Lyons, K. E., Pahwa, R., Burns, J., and Kemper, S. (2011). Understanding verbal fluency in healthy aging, alzheimer's disease, and parkinson's disease. *Neuropsychology*, 25(2):210.

Melamud, O., Goldberger, J., and Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pages 51–61.

Midi, İ., Doğan, M., Pata, Y. S., Koçak, İ., Mollahasanoğlu, A., and Tuncer, N. (2011). The effects of verbal reaction time in Alzheimer's disease. *Laryngoscope*, 121(7):1495–1503.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2017). Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Millar, R. (2002). Towards a science curriculum for public understanding. *Teaching science in secondary schools*, pages 113–128.

Miller, G. A. (1995). *WordNet : A Lexical Database for English*.

Milner, B. (1964). Some effects of frontal lobectomy in man. *The frontal granular cortex and behavior.*, pages 313–334.

Mirandez, R. M., Aprahamian, I., Talib, L. L., Forlenza, O. V., and Radanovic, M. (2017). Multiple category verbal fluency in mild cognitive impairment and correlation with csf biomarkers for alzheimer's disease. *International psychogeriatrics*, 29(6):949–958.

Mironets, S., Deviaterikova, A., Glebova, E., and Kasatkin, V. (2023). Phonological verbal fluency test for russian-speaking children. *Folia Phoniatrica et Logopaedica*, pages 1–8.

Mitchell, R. L. and Ross, E. D. (2013). Attitudinal prosody: What we know and directions for future study. *Neuroscience & Biobehavioral Reviews*, 37(3):471–479.

Mok, E. H. L., Lam, L. C. W., and Chiu, H. F. K. (2004). Category verbal fluency test performance in chinese elderly with alzheimer's disease. *Dementia and geriatric cognitive disorders*, 18(2):120–124.

Monsch, A. U., Bondi, M. W., Butters, N., Salmon, D. P., Katzman, R., and Thal, L. J. (1992). Comparisons of verbal fluency tasks in the detection of dementia of the alzheimer type. *Archives of neurology*, 49(12):1253–1258.

Monsch, A. U., Bondi, M. W., Butters, N., Salmon, D. P., Katzman, R., and Thal, L. J. (2020). Comparisons of Verbal Fluency Tasks in the of Dementia of the Alzheimer. *Archives of neurology*.

Moraes, A. L., Guimarães, L. S., Joanette, Y., Parente, M. A. d. M. P., Fonseca, R. P., and Almeida, R. M. M. d. (2013). Effect of aging, education, reading and writing, semantic processing and depression symptoms on verbal fluency. *Psicologia: Reflexão e Crítica*, 26:680–690.

Morris, J. C., Heyman, A., Mohs, R. C., Hughes, J., van Belle, G., Fillenbaum, G., Mellits, E., and Clark, C. (1989). The consortium to establish a registry for alzheimer's disease (cerad). part i. clinical and neuropsychological assessment of alzheimer's disease. *Neurology*, 39(9):1159–1165.

Morrison, G. E., Simone, C. M., Ng, N. F., and Hardy, J. L. (2015). Reliability and validity of the neurocognitive performance test, a web-based neuropsychological assessment. *Frontiers in psychology*, page 1652.

Mortensen, L., Meyer, A. S., and Humphreys, G. W. (2006). Age-related effects on speech production: A review. *Language and Cognitive Processes*, 21(1-3):238–290.

Mueller Gathercole, V. C., Thomas, E. M., Jones, L., Guasch, N. V., Young, N., and Hughes, E. K. (2010). Cognitive effects of bilingualism: digging deeper for the contributions of language dominance, linguistic knowledge, socio-economic status and cognitive abilities. *International Journal of Bilingual Education and Bilingualism*, 13(5):617–664.

Mutlu, E. (2018). *Düşünce ve dil bozukluğu ölçeğinin Türkçe uyarlaması ve şizofrenide düşünce bozukluğunun hastalık şiddeti,bilişsel işlevler, genel ve sosyal işlevsellik ve yaşam kalitesi ile ilişkisi.* PhD thesis, Hacettepe University.

Mutlu, E., Abaoğlu, H., Barışkın, E., Gürel, Ş. C., Ertuğrul, A., Yazıcı, M. K., Akı, E., and Yağcıoğlu, A. E. A. (2021). The cognitive aspect of formal thought disorder and its relationship with global social functioning and the quality of life in schizophrenia. *Social Psychiatry and Psychiatric Epidemiology*, 56:1399–1410.

Namey, E., Guest, G., O'Regan, A., Godwin, C. L., Taylor, J., and Martinez, A. (2020). How does mode of qualitative data collection affect data and cost? findings from a quasi-experimental study. *Field methods*, 32(1):58–74.

Naseem, U., Razzak, I., Khan, S. K., and Prasad, M. (2021). A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–35.

National Brain Tumor Society, U. (2023). Areas of the Brain.

Neelima, A. and Mehrotra, S. (2023). A comprehensive review on word embedding techniques. In *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, pages 538–543. IEEE.

Nielsen, T. R., Vogel, A., and Waldemar, G. (2012). Comparison of performance on three neuropsychological tests in healthy Turkish immigrants and Danish elderly. *International Psychogeriatrics*, 24(9):1515–1521.

Nielsen, T. R. and Waldemar, G. (2016). Effects of literacy on semantic verbal fluency in an immigrant population. *Aging, Neuropsychology, and Cognition*, 23(5):578–590.

Nighojkar, A., Khlyzova, A., and Licato, J. (2022). Cognitive modeling of semantic fluency using transformers. *arXiv preprint arXiv:2208.09719*.

Nocera, J., Crosson, B., Mammino, K., McGregor, K. M., et al. (2017). Changes in cortical activation patterns in language areas following an aerobic exercise intervention in older adults. *Neural Plasticity*, 2017.

Nocera, J. R., Mammino, K., Kommula, Y., Wharton, W., Crosson, B., and McGregor, K. M. (2020). Effects of combined aerobic exercise and cognitive training on verbal fluency in older adults. *Gerontology and Geriatric Medicine*, 6:2333721419896884.

Nogueira, D. S., Reis, E. A., and Vieira, A. (2016). Verbal fluency tasks: Effects of age, gender, and education. *Folia phoniatrica et logopaedica*, 68(3):124–133.

Nosheny, R. L., Flennkiken, D., Insel, P. S., Finley, S., Mackin, S., Camacho, M., Truran-Sacrey, D., Maruff, P., and Weiner, M. W. (2015). O1-10-06: Internet-based recruitment of subjects for prodromal and secondary prevention alzheimer's disease trials using the brain health registry. *Alzheimer's & Dementia*, 11(7S_Part_3):P156–P156.

Noyan, H. (2011). *Cognitive dysfunction related to subtypes of schizotypy in healty first-degree relatives of patient with schizophrenia*. PhD thesis, Istanbul University.

Ober, B. A., Dronkers, N. F., Koss, E., Delis, D. C., and Friedland, R. P. (1986). Retrieval from semantic memory in alzheimer-type dementia. *Journal of clinical and experimental neuropsychology*, 8(1):75–92.

Obeso, I., Casabona, E., Bringas, M. L., Álvarez, L., and Jahanshahi, M. (2012). Semantic and phonemic verbal fluency in parkinson's disease: Influence of clinical and demographic variables. *Behavioural neurology*, 25(2):111–118.

O'connor, P. (1990). Normative data: their definition, interpretation, and importance for primary care physicians. *Family medicine*, 22(4):307–311.

Oh, S. J., Sung, J. E., Choi, S. J., and Jeong, J. H. (2019). Clustering and switching patterns in semantic fluency and their relationship to working memory in mild cognitive impairment. *Dementia and neurocognitive disorders*, 18(2):47–61.

Öhman, H., Savikko, N., Strandberg, T. E., Kautiainen, H., Raivio, M. M., Laakkonen, M.-L., Tilvis, R., and Pitkälä, K. H. (2016). Effects of exercise on cognition: the finnish alzheimer disease exercise trial: a randomized, controlled trial. *Journal of the American Geriatrics Society*, 64(4):731–738.

Olabarrieta-Landa, L., Torre, E. L., López-Mugartza, J. C., Bialystok, E., and Arango-Lasprilla, J. C. (2017). Verbal fluency tests: Developing a new model of administration and scoring for spanish language. *NeuroRehabilitation*, 41(2):539–565.

Önder, B. (2019). *Neurocognitive flexibility, perfectionism and obsessive beliefs in obsessive-compulsive disorder*. PhD thesis, Dokuz Eylül University.

Ostrosky-Solis, F., Gutierrez, A. L., Flores, M. R., and Ardila, A. (2007). Same or different? semantic verbal fluency across spanish-speakers from different countries. *Archives of clinical neuropsychology*, 22(3):367–377.

Özcan, H. (2010). *Neuropsychological, Electrophysiological and Neurological Impairments in Patients with Obsessive Compulsive Disorder and Their Healthy Siblings.* PhD thesis, Hacettepe University.

Özcan, H., Özer, S., and Yağcıoğlu, S. (2016). Neuropsychological, electrophysiological and neurological impairments in patients with obsessive compulsive disorder, their healthy siblings and healthy controls: Identifying potential endophenotype(s). *Psychiatry Research*, 240:110–117.

Özçelik-Eroğlu, E. (2012). *The influence of clozapine on diffusion tensor imaging measures in patients with schizophrenia*. PhD thesis, Hacettepe University.

Özçelik-Eroğlu, E., Ertuğrul, A., Oguz, K. K., Has, A. C., Karahan, S., and Yazıcı, M. K. (2014). Effect of clozapine on white matter integrity in patients with schizophrenia: A diffusion tensor imaging study. *Psychiatry Research - Neuroimaging*, 223(3):226–235.

Özdemir, S. (2015). *Assesment of Semantic and Action Fluency skills among school age children ranging in age from 15 to 17 and Adults above 18 years age*. PhD thesis, Anadolu University.

Özdemir, S. and Tunçer, A. (2021). Verbal Fluency: An Investigation of Time Variable Among Elderly People. *Clinical and Experimental Health Sciences*, 11(1).

Özkul, Ç., Güçlü-Gunduz, A., Eldemir, K., Apaydin, Y., Gülsen, Ç., Yazici, G., Söke, F., and Irkec, C. (2021). Dual-task cost and related clinical features in patients with multiple sclerosis. *Motor Control*, 25(2):211–233.

Paap, K. R., Myuz, H. A., Anders, R. T., Bockelman, M. F., Mikulinsky, R., and Sawi, O. M. (2017). No compelling evidence for a bilingual advantage in switching or that frequent language switching reduces switch cost. *Journal of Cognitive Psychology*, 29(2):89–112.

Pagliarin, K. C., Fernandes, E. G., Muller, M. D., Portalete, C. R., Fonseca, R. P., and Altmann, R. F. (2021). Clustering and switching in verbal fluency: a comparison between control and individuals with brain damage. In *CoDAS*, volume 34, page e20200365. SciELO Brasil.

Pakhomov, S. V. and Hemmy, L. S. (2014). A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the nun study. *Cortex*, 55:97–106.

Pakhomov, S. V., Hemmy, L. S., and Lim, K. O. (2012). Automated semantic indices related to cognitive function and rate of cognitive decline. *Neuropsychologia*, 50(9):2165–2175.

Pakhomov, S. V., Marino, S. E., Banks, S., and Bernick, C. (2015). Using automatic speech recognition to assess spoken responses to cognitive tests of semantic verbal fluency. *Speech communication*, 75:14–26.

Patra, A., Bose, A., and Marinis, T. (2020). Performance difference in verbal fluency in bilingual and monolingual speakers. *Bilingualism: Language and Cognition*, 23(1):204–218.

Patterson, J., Kreutzer, J. S., DeLuca, J., and Caplan, B., editors (2011). *Verbal Fluency*, pages 2603–2606. Springer New York, New York, NY.

Paula, F., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). Similarity Measures for the Detection of Clinical Conditions with Verbal Fluency Tasks. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 231–235.

Pekkala, S., Albert, M. L., Spiro III, A., and Erkinjuntti, T. (2008). Perseveration in alzheimer's disease. *Dementia and geriatric cognitive disorders*, 25(2):109–114.

Pekkala, S., Goral, M., Hyun, J., Obler, L. K., Erkinjuntti, T., and Albert, M. L. (2009). Semantic verbal fluency in two contrasting languages. *Clinical linguistics & phonetics*, 23(6):431–445.

Pendlebury, S. T., Welch, S. J., Cuthbertson, F. C., Mariz, J., Mehta, Z., and Rothwell, P. M. (2013). Telephone assessment of cognition after transient ischemic attack and stroke: modified telephone interview of cognitive status and telephone montreal cognitive assessment versus face-to-face montreal cognitive assessment and neuropsychological battery. *Stroke*, 44(1):227–229.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pereira, A. H., Gonçalves, A. B., Holz, M., Gonçalves, H. A., Kochhann, R., Joanette, Y., Zimmermann, N., and Fonseca, R. P. (2018). Influence of age and education on the processing of clustering and switching in verbal fluency tasks. *Dementia & neuropsychologia*, 12:360–367.

Perlmutter, M. (1978). What is memory aging the aging of? *Developmental Psychology*, 14(4):330.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (1802). Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:1802.05365*.

Pham, D.-H. and Le, A.-C. (2018). Exploiting multiple word embeddings and one-hot character vectors for aspect-based sentiment analysis. *International Journal of Approximate Reasoning*, 103:1–10.

Piatt, A. L., Fields, J. A., Paolo, A. M., Koller, W. C., and Tröster, A. I. (1999). Lexical, semantic, and action verbal fluency in parkinson's disease with and without dementia. *Journal of Clinical and Experimental Neuropsychology*, 21(4):435–443.

Piatt, A. L., Fields, J. A., Paolo, A. M., Koller, W. C., Tröster, A. I., Piatt, A. L., Fields, J. A., Paolo, A. M., Koller, W. C., Piatt, A. L., Fields, J. A., Paolo, A. M., Koller, W. C., and Tröster, A. I. (2010). Lexical , Semantic , and Action Verbal Fluency in Parkinson ' s Disease with and without Dementia Lexical , Semantic , and Action

Verbal Fluency in Parkinson ' s Disease with and without Dementia *. *Journal of Clinical and Experimental Neuropsychology*, 3395.

Pietrowicz, M., Agurto, C., Norel, R., Eyigöz, E., Cecchi, G. A., Bilgrami, Z. R., and Corcoran, C. (2019). A new approach for automating analysis of responses on verbal fluency tests from subjects at-risk for schizophrenia. In *INTERSPEECH*, pages 3028–3032.

Porter, M. (2001). A language for stemming algorithms.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Portocarrero, J. S., Burright, R. G., and Donovick, P. J. (2007). Vocabulary and verbal fluency of bilingual and monolingual college students. *Archives of Clinical Neuropsychology*, 22(3):415–422.

Prud'hommeaux, E., Santen, J. V., and Gliner, D. (2017). Vector space models for evaluating semantic fluency in autism. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2:32–37.

Prud'hommeaux, E., Van Santen, J., and Gliner, D. (2017). Vector space models for evaluating semantic fluency in autism. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 32–37.

Qader, W. A., Ameen, M. M., and Ahmed, B. I. (2019). An overview of bag of words; importance, implementation, applications, and challenges. In *2019 international engineering conference (IEC)*, pages 200–204. IEEE.

Qaiser, S. and Ali, R. (2018). Text mining: use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1):25–29.

Qualtrics, D. C. (2005). Qualtrics.

Quaranta, D., Piccininni, C., Caprara, A., Malandrino, A., Gainotti, G., and Marra, C. (2019). Semantic relations in a categorical verbal fluency test: an exploratory investigation in mild cognitive impairment. *Frontiers in psychology*, 10:2797.

Raboutet, C., Sauzéon, H., Corsini, M.-M., Rodrigues, J., Langevin, S., and N'kaoua, B. (2010). Performance on a semantic verbal fluency task across time: Dissociation

between clustering, switching, and categorical exploitation processes. *Journal of Clinical and Experimental Neuropsychology*, 32(3):268–280.

Radanovic, M., Diniz, B. S., Mirandez, R. M., da Silva Novaretti, T. M., Flacks, M. K., Yassuda, M. S., and Forlenza, O. V. (2009). Verbal fluency in the detection of mild cognitive impairment and alzheimer's disease among brazilian portuguese speakers: the influence of education. *International Psychogeriatrics*, 21(6):1081–1087.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. *preprint*.

Rahman, Q., Abrahams, S., and Wilson, G. D. (2003). Sexual-orientation-related differences in verbal fluency. *Neuropsychology*, 17(2):240.

Rahman, Q., Andersson, D., and Govier, E. (2005). A specific sexual orientation-related difference in navigation strategy. *Behavioral neuroscience*, 119(1):311.

Rahman, Q. and Wilson, G. D. (2003). Large sexual-orientation-related differences in performance on mental rotation and judgement of line orientation tasks. *Neuropsychology*, 17(1):25.

Raina, P. S., Wolfson, C., Kirkland, S. A., Griffith, L. E., Oremus, M., Patterson, C., Tuokko, H., Penning, M., Balion, C. M., Hogan, D., et al. (2009). The canadian longitudinal study on aging (clsa). *Canadian Journal on Aging/La Revue canadienne du vieillissement*, 28(3):221–229.

Ralph, M. A. L., Jefferies, E., Patterson, K., and Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature reviews neuroscience*, 18(1):42–55.

Randolph, C., Braun, A. R., Goldberg, T. E., and Chase, T. N. (1993). Semantic fluency in alzheimer's, parkinson's, and huntington's disease: Dissociation of storage and retrieval failures. *Neuropsychology*, 7(1):82.

Raoux, N., Amieva, H., Le Goff, M., Auriacombe, S., Carcaillon, L., Letenneur, L., and Dartigues, J.-F. (2008). Clustering and switching processes in semantic verbal fluency in the course of alzheimer's disease subjects: Results from the paquid longitudinal study. *Cortex*, 44(9):1188–1196.

Raskin, S. A. and Rearick, E. (1996a). Verbal fluency in individuals with mild traumatic brain injury. *Neuropsychology*, 10(3):416.

Raskin, S. A. and Rearick, E. (1996b). Verbal Fluency in Individuals With Mild Traumatic Brain Injury. *Neuropsychology*, 10(3):416–422.

Raskin, S. A., Sliwinski, M., and Borod, J. C. (1992a). Clustering strategies on tasks of verbal fluency in parkinson's disease. *Neuropsychologia*, 30(1):95–99.

Raskin, S. A., Sliwinski, M., and Borod, J. C. (1992b). Clustering strategies on tasks of verbal fluency in Parkinson's disease. *Neuropsychologia*, 30(1):95–99.

Ratcliff, G., Ganguli, M., Chandra, V., Sharma, S., Belle, S., Seaberg, E., and Pandav, R. (1998). Effects of literacy and education on measures of word fluency. *Brain and Language*, 61(1):115–122.

Raucher-Chéné, D., Achim, A. M., Kaladjian, A., and Besche-Richard, C. (2017). Verbal fluency in bipolar disorders: A systematic review and meta-analysis. *Journal of affective disorders*, 207:359–366.

Rehurek, R. and Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Richardson, W. S., Wilson, M. C., Nishikawa, J., and Hayward, R. S. (1995). The well-built clinical question: a key to evidence-based decisions. *ACP journal club*, 123(3):7–9.

Robert, P. H., Lafont, V., Medecin, I., Berthet, L., Thauby, S., Baudu, C., and Darcourt, G. (1998). Clustering and switching strategies in verbal fluency tasks: comparison between schizophrenics and healthy adults. *Journal of the International Neuropsychological Society*, 4(6):539–546.

Robinson, G., Shallice, T., Bozzali, M., and Cipolotti, L. (2012). The differing roles of the frontal cortex in fluency tests. *Brain*, 135(7):2202–2214.

Rodríguez-aranda, C. and Martinussen, M. (2006). Age-Related Differences in Performance of Phonemic Verbal Fluency Measured by Controlled Oral Word Association Task (COWAT): A Meta-Analytic Study. *Developmental Neuropsychology*, 30(2):697–717.

Rodríguez-Lorenzana, A., Benito-Sánchez, I., Adana-Díaz, L., Paz, C. P., Yacelga Ponce, T., Rivera, D., and Arango-Lasprilla, J. C. (2020). Normative data for test of verbal fluency and naming on ecuadorian adult population. *Frontiers in Psychology*, 11:830.

Rong, X. (2014). word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.

Rose, M. R. (1991). *Evolutionary biology of aging*. Oxford University Press, USA.

Rosenstein, M., Foltz, P., Vaskinn, A., and Elvevåg, B. (2015). Practical issues in developing semantic frameworks for the analysis of verbal fluency data: A norwegian data case study. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From linguistic signal to clinical reality*, pages 124–133.

Rosselli, M., Ardila, A., Araujo, K., Weekes, V. A., Caracciolo, V., Padilla, M., and Ostrosky-Solí, F. (2000a). Verbal fluency and repetition skills in healthy older spanish-english bilinguals. *Applied neuropsychology*, 7(1):17–24.

Rosselli, M., Ardila, A., Moreno, S., Standish, V., Arango-Lasprilla, J. C., Tirado, V., Ossa, J., Goate, A. M., Kosik, K. S., and Lopera, F. (2000b). Cognitive decline in patients with familial alzheimer's disease associated with e280a presenilin-1 mutation: a longitudinal study. *Journal of Clinical and Experimental Neuropsychology*, 22(4):483–495.

Rosselli, M., Ardila, A., Salvatierra, J., Marquez, M., LUIS, M., and Weekes, V. A. (2002). A cross-linguistic comparison of verbal fluency tests. *International Journal of Neuroscience*, 112(6):759–776.

Rosselli, M., Tappen, R., Williams, C., Salvatierra, J., and Zoller, Y. (2009). Level of education and category fluency task among spanish speaking elders: number of words, clustering, and switching strategies. *Aging, Neuropsychology, and Cognition*, 16(6):721–744.

Rosser, A. and Hodges, J. R. (1994). Initial letter and semantic category fluency in alzheimer's disease, huntington's disease, and progressive supranuclear palsy. *Journal of Neurology, Neurosurgery & Psychiatry*, 57(11):1389–1394.

Ruff, R., Light, R., Parker, S., and Levin, H. (1996). Benton controlled oral word association test: reliability and updated norms. *Archives of clinical neuropsychology*, 11(4):329–338.

Ryan, J. O., Pakhomov, S., Marino, S., Bernick, C., and Banks, S. (2013). Computerized analysis of a verbal fluency test. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 884–889.

Rybatzki, V. (2020). The altaic languages: Tungusic, mongolic, turkic. In *The Oxford Guide to the Transeurasian Languages*, pages 22–28. Oxford University Press.

Sahin, M. C. (2023). *Cognitive Functions And Posture Analysis Of University Students Who Play Video Games, Athletes And Musicians*. PhD thesis, Bursa Uludağ University.

Sahin, R. (2022). *Evaluation of executive functions and attention in epilepsy patients with multimethod measurement techniques*. PhD thesis, Akdeniz University.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Sandoval, T. C., Gollan, T. H., Ferreira, V. S., and Salmon, D. P. (2010). What causes the bilingual disadvantage in verbal fluency? the dual-task analogy. *Bilingualism: Language and Cognition*, 13(2):231–252.

Sandson, J. and Albert, M. L. (1984). Varieties of perseveration. *Neuropsychologia*, 22(6):715–732.

Sandson, J. and Albert, M. L. (1987). Perseveration in behavioral neurology. *Neurology*, 37(11):1736–1736.

Saranpää, A. M., Kivisaari, S. L., Salmelin, R., and Krumm, S. (2022). Moving in semantic space in prodromal and very early alzheimer's disease: An item-level characterization of the semantic fluency task. *Frontiers in Psychology*, 13:777656.

Sarma, D., Mittra, T., and Hossain, M. S. (2021). Personalized book recommendation system using machine learning algorithm. *International Journal of Advanced Computer Science and Applications*, 12(1).

Scheuringer, A. and Pletzer, B. (2017). Sex differences and menstrual cycle dependent changes in cognitive strategies during spatial navigation and verbal fluency. *Frontiers in Psychology*, 8:381.

Scheuringer, A., Wittig, R., and Pletzer, B. (2017). Sex differences in verbal fluency: The role of strategies and instructions. *Cognitive processing*, 18:407–417.

Şentürk, T. (2019). *Turkish Normative Data Of Semantic And Phonemic Verbal Fluency Tests*. PhD thesis, Dokuz Eylül University.

Sezikli, S. (2014). *Comparsion of frontal lob cogntive functions of JME patients who have asymmetrical eeg results and those who have not*. PhD thesis, Mashar Osman Ruh Sağlığı ve Sinir Hastalıklari Eğitim ve Araştırma Hastanesi.

Sezikli, S., Pulat, T. A., Tekin, B., Ak, P. D., Keskinkiliç, C., and Atakli, D. (2018). Frontal lobe cognitive functions and electroencephalographic features in juvenile myoclonic epilepsy. *Epilepsy and Behavior*, 86:102–107.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Shao, Z., Janse, E., Visser, K., and Meyer, A. S. (2014). What do verbal fluency tasks measure? predictors of verbal fluency performance in older adults. *Frontiers in psychology*, 5:772.

Singhal, A., Salton, G., Mitra, M., and Buckley, C. (1996). Document length normalization. *Information Processing & Management*, 32(5):619–633.

Sokołowski, A., Tyburski, E., Sołtys, A., and Karabanowicz, E. (2020). Sex differences in verbal fluency among young adults. *Advances in Cognitive Psychology*, 16(2):92.

Soriano, F., Fumagalli, J., Shalóm, D., Carden, J., Borovinsky, G., Manes, F., and Martínez-Cuitiño, M. (2015). Sex differences in a semantic fluency task. *East European Journal of Psycholinguistics*, 2(1):134–140.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Staffaroni, A. M., Tsoy, E., Taylor, J., Boxer, A. L., and Possin, K. L. (2020). Digital cognitive assessments for dementia: Digital assessments may enhance the efficiency of evaluations in neurology and other clinics. *Practical Neurology (Fort Washington, Pa.)*, 2020:24.

Sternin, A., Burns, A., and Owen, A. M. (2019). Thirty-five years of computerized cognitive assessment of aging—where are we now? *Diagnostics*, 9(3):114.

Stokholm, J., Jørgensen, K., and Vogel, A. (2013). Performances on five verbal fluency tests in a healthy, elderly danish sample. *Aging, Neuropsychology, and Cognition*, 20(1):22–33.

Strauss, E., Sherman, E. M., and Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. American chemical society.

Stuss, D. T., Alexander, M. P., Hamer, L., Palumbo, C., Dempster, R., Binns, M., Levine, B., and Izukawa, D. (1998). The effects of focal anterior and posterior brain lesions on verbal fluency. *Journal of the International Neuropsychological Society*, 4(3):265–278.

Sumiyoshi, C., Ertugrul, A., Yağcıoğlu, A. E. A., Roy, A., Jayathilake, K., Milby, A., Meltzer, H. Y., and Sumiyoshi, T. (2014a). Language-dependent performance on the letter fluency task in patients with schizophrenia. *Schizophrenia research*, 152(2-3):421–429.

Sumiyoshi, C., Ertuğrul, A., Yağcıoğlu, A. E. A., Roy, A., Jayathilake, K., Milby, A., Meltzer, H. Y., and Sumiyoshi, T. (2014b). Language-dependent performance on the letter fluency task in patients with schizophrenia. *Schizophrenia Research*, 152(2-3):421–429.

Takács, Á., Kóbor, A., Tárnok, Z., and Csépe, V. (2014). Verbal fluency in children with adhd: strategy using and temporal properties. *Child Neuropsychology*, 20(4):415–429.

Talas, M. S. (2009). *Examination of the relationships between verbal fluency, magical ideation and motor asymmetry in healthy young adults*. PhD thesis, Ankara University.

Taler, V., Johns, B. T., and Jones, M. N. (2020). A large-scale semantic analysis of verbal fluency across the aging spectrum: Data from the canadian longitudinal study on aging. *The Journals of Gerontology: Series B*, 75(9):e221–e230.

Taler, V., Johns, B. T., Young, K., Sheppard, C., and Jones, M. N. (2013). A computational analysis of semantic structure in bilingual verbal fluency performance. *Journal of Memory and Language*, 69(4):607–618.

Thurstone, L. L. and Thurstone, T. G. (1938). *Primary mental abilities*, volume 119. University of Chicago Press Chicago.

Tombaugh, T. N., Kozak, J., and Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: Fas and animal naming. *Archives of clinical neuropsychology*, 14(2):167–177.

Tomer, R. and Levin, B. E. (1993). Differential effects of aging on two verbal fluency tasks. *Perceptual and motor skills*, 76(2):465–466.

Töret, Z. (2019). *The comparison of the effects of gesture use on the lexical access of adults with visual impairments and sighted adults*. PhD thesis, Gazi University.

Töret, Z. and Özdemir, S. (2021). Comparison of the Effects of Gesture Usage on the Lexical Transportation Process of Visually Impaired and Sighted Adults. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi*, 22(2).

Tosto, G., Gasparini, M., Lenzi, G. L., and Bruno, G. (2011). Prosodic impairment in alzheimer's disease: assessment and clinical relevance. *The Journal of neuropsychiatry and clinical neurosciences*, 23(2):E21–E23.

TRNC, T. R. o. N. C. S. I. (2011). Population and Demography Bulletin.

Tröger, J., Linz, N., König, A., Robert, P., and Alexandersson, J. (2018). Telephone-based dementia screening i: automated semantic verbal fluency assessment. In *Proceedings of the 12th EAI International Conference on pervasive computing technologies for healthcare*, pages 59–66.

Tröger, J., Linz, N., König, A., Robert, P., Alexandersson, J., Peter, J., and Kray, J. (2019). Exploitation vs. exploration—computational temporal and semantic analysis explains semantic verbal fluency impairment in alzheimer's disease. *Neuropsychologia*, 131:53–61.

Tröster, A. I., Fields, J. A., Testa, J. A., Paul, R. H., Blanco, C. R., Hames, K. A., Salmon, D. P., and Beatty, W. W. (1998). Cortical and subcortical influences on clustering and switching in the performance of verbal fluency tasks. *Neuropsychologia*, 36(4):295–304.

Tröster, A. I., Salmon, D. P., McCullough, D., and Butters, N. (1989). A comparison of the category fluency deficits associated with alzheimer's and huntington's disease. *Brain and Language*, 37(3):500–513.

Troyer, A. K. (2000). Normative data for clustering and switching on verbal fluency tasks. *Journal of clinical and experimental neuropsychology*, 22(3):370–378.

Troyer, A. K., Moscovitch, M., and Winocur, G. (1997). Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology*, 11(1):138–146.

Troyer, A. K., Moscovitch, M., Winocur, G., Alexander, M. P., and Stuss, D. (1998a). Clustering and switching on verbal fluency: The effects of focal frontal-and temporal-lobe lesions. *Neuropsychologia*, 36(6):499–504.

Troyer, A. K., Moscovitch, M., Winocur, G., Leach, L., and Freedman, M. (1998b). Clustering and switching on verbal fluency tests in alzheimer's and parkinson's disease. *Journal of the International Neuropsychological Society*, 4(2):137–143.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., and Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98.

TUIK, T. S. I. (2022). World Population Day, 2022.

Tumaç, A. (1997). Effects of age and education to performance in some frontal lobe tests in normal subjects. Master's thesis, Istanbul University.

Tuncer, A. M. (2012). *Verbal Fluency Performance of Turkish Speaking Adults*. PhD thesis, Anadolu University.

Turkish Ministry of Foreign Affairs, T. (2022). Turkish Citizens Living Abroad.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Türkmen, H., Dikenelli, O., Eraslan, C., Çallı, M. C., and Özbek, S. S. (2023). Harnessing the power of bert in the turkish clinical domain: Pretraining approaches for limited data scenarios.

Ünlü, C., Yapıcı, N., İzgi, F. C., Kudisoğlu, T., Ünlü, C., and Aykaç, Z. (2013). Effects of N-acetylcysteine on neurocognitive functions after coronary artery bypass graft surgery. *Türk Göğüs Kalp Damar Cerrahisi Dergisi*, 21(2).

Unsworth, N., Spillers, G. J., and Brewer, G. A. (2011). Variation in verbal fluency: A latent variable analysis of clustering, switching, and overall performance. *Quarterly Journal of Experimental Psychology*, 64(3):447–466.

Upadhayay, N. and Guragain, S. (2014). Comparison of cognitive functions between male and female medical students: a pilot study. *Journal of clinical and diagnostic research: JCDR*, 8(6):BC12.

Uslu, E. (2012). *The investigation of bipolar subtypes which have been proposed to have different genetic etiology with respect to nerucognitive functions and temperamental traits: A controlled study*. PhD thesis, Hacettepe University.

Uzgan, B. Ö., Oktay, M. T., Aykaç, C., Ermiş, Ç., and Alkın, T. (2021). Neurocognitive flexibility, perfectionism, obsessive beliefs in patients with obsessive compulsive disorder. *Klinik Psikiyatri Dergisi*, 24(4).

Van Der Elst, W., Van Boxtel, M. P., Van Breukelen, G. J., and Jolles, J. (2006a). Normative data for the animal, profession and letter m naming verbal fluency tests for dutch speaking participants and the effects of age, education, and sex. *Journal of the International Neuropsychological Society*, 12(1):80–89.

Van Der Elst, W. I. M., Van Boxtel, M. P., Van Breukelen, G. J., and Jolles, J. (2006b). Normative data for the Animal , Profession and Letter M Naming verbal fluency tests for Dutch speaking participants and the effects of age , education , and sex. *Journal of the International Neuropsychological Society*, 12(1):80–89.

Van Wagenen, A., Driskell, J., and Bradford, J. (2013). "i'm still raring to go": Successful aging among lesbian, gay, bisexual, and transgender older adults. *Journal of aging studies*, 27(1):1–14.

Venekoski, V. and Vankka, J. (2017). Finnish resources for evaluating language model semantics. In *Proceedings of the 21st Nordic conference on computational linguistics*, pages 231–236.

Vermeent, S., Dotsch, R., Schmand, B., Klaming, L., Miller, J. B., and Van Elswijk, G. (2020). Evidence of validity for a newly developed digital cognitive test battery. *Frontiers in Psychology*, 11:770.

Vidal, A., Puig, O., Boget, T., and Salamero, M. (2006). Gender differences in cognitive functions and influence of sex hormones. *Actas Esp Psiquiatr*, 34(6):408–415.

Vilkki, J. and Holst, P. (1994). Speed and flexibility on word fluency tasks after focal brain lesions. *Neuropsychologia*, 32(10):1257–1262.

Villalobos, D., Torres-Simón, L., Pacios, J., Paúl, N., and Del Río, D. (2022). A systematic review of normative data for verbal fluency test in different languages. *Neuropsychology Review*, pages 1–32.

Vlachos, F., Andreou, G., and Andreou, E. (2003). Biological and environmental influences in visuospatial abilities. *Learning and Individual Differences*, 13(4):339–347.

Voppel, A., de Boer, J., Brederoo, S., Schnack, H., and Sommer, I. (2021). Quantified language connectedness in schizophrenia-spectrum disorders. *Psychiatry Research*, 304:114130.

Wall, K. J., Cumming, T. B., and Copland, D. A. (2017). Determining the association between language and cognitive tests in poststroke aphasia. *Frontiers in neurology*, 8:149.

Weiner, L., Guidi, A., Doignon-Camus, N., Giersch, A., Bertschy, G., and Vanello, N. (2021). Vocal features obtained through automated methods in verbal fluency tasks can aid the identification of mixed episodes in bipolar disorder. *Translational Psychiatry*, 11(1):415.

Weiner, M. W., Nosheny, R., Camacho, M., Truran-Sacrey, D., Mackin, R. S., Flenniken, D., Ulbricht, A., Insel, P., Finley, S., Fockler, J., et al. (2018). The brain health registry: an internet-based platform for recruitment, assessment, and longitudinal monitoring of participants for neuroscience studies. *Alzheimer's & Dementia*, 14(8):1063–1076.

Weiss, E. M., Ragland, J. D., Brensinger, C. M., Bilker, W. B., Deisenhammer, E. A., and Delazer, M. (2006). Sex differences in clustering and switching in verbal fluency tasks. *Journal of the International Neuropsychological Society*, 12(4):502–509.

Welford, P., Östh, J., Hoy, S., L Rossell, S., Pascoe, M., Diwan, V., and Hallgren, M. (2023). Effects of yoga and aerobic exercise on verbal fluency in physically inactive older adults: Randomized controlled trial (fitforage). *Clinical Interventions in Aging*, pages 533–545.

Wolters, M. K., Kim, N., Kim, J.-H., MacPherson, S. E., and Park, J. C. (2016). Prosodic and linguistic analysis of semantic fluency data: A window into speech production and cognition. In *Interspeech*, pages 2085–2089. San Francisco, CA.

Wolters, M. K., Macpherson, S. E., You, J., Jin, R., Baek, S.-C., and Park, J. C. (2015). A new measure of clustering and switching based on bigrams.

Wood, T. (2006). The united states of america: Mathematics education reform and the united states educational system. In *Mathematics Classrooms in Twelve Countries*, pages 373–375. Brill.

Woods, D. L., Wyma, J. M., Herron, T. J., and Yund, E. W. (2016a). Computerized Analysis of Verbal Fluency : Normative Data and the Effects of Computerized Analysis of Verbal Fluency : Normative Data and the Effects of Repeated Testing , Simulated Malingering , and Traumatic Brain Injury. *PloS one*, 11(12).

Woods, D. L., Wyma, J. M., Herron, T. J., and Yund, E. W. (2016b). Computerized analysis of verbal fluency: Normative data and the effects of repeated testing, simulated malingering, and traumatic brain injury. *PloS one*, 11(12):e0166439.

Wu, L., Hoi, S. C., and Yu, N. (2010). Semantics-preserving bag-of-words models and applications. *IEEE Transactions on Image Processing*, 19(7):1908–1920.

Yan, Y.-j., Lin, R., Zhou, Y., Luo, Y.-t., Cai, Z.-z., Zhu, K.-y., and Li, H. (2021). Effects of expressive arts therapy in older adults with mild cognitive impairment: A pilot study. *Geriatric Nursing*, 42(1):129–136.

Yavuz-Demiray, D. (2011). *Assessment of cognitive impairment and depression in patients with Essential tremor, patients with Essential tremor and Parkinson's disease concomitance and Parkinson's disease*. PhD thesis, Istanbul University.

Yazici, M. (2019). *Neuropsychological Evaluation Of The Population Of Eastern And Southeastern Anatolia On Executive Functions And Complex Attention*. PhD thesis, Ondokuz Mayis University.

Yeniçeri, F. E. (2019). *The Effect Of Balance Exercises With Cognitive Task On Motor And Cognitive Functions In Healthy Young Adults*. PhD thesis, Medipol University.

Yildirim, E. and Ogel-Balaban, H. (2021). Cognitive functions among healthy older adults using online social networking. *Applied Neuropsychology: Adult*, 30(4):401–408.

Yilmaz, T. (2014). *Comparison of adults diagnosed with attention deficit and hyperactivity disorder with healthy controls on neurocognitive functions, impulsivity and theory of mind*. PhD thesis, Maltepe University.

Yılmaz, T., Karaş, H., and Tan, D. (2020). Relationship between theory of mind, impulsivity and cognitive functions in adult attention deficit and hyperactivity disorder. *Anadolu Psikiyatri Dergisi*, 21(2):149–157.

Young, R., Tischler, V., Hulbert, S., and Camic, P. M. (2015). The impact of viewing and making art on verbal fluency and memory in people with dementia in an art gallery setting. *Psychology of Aesthetics, Creativity, and the Arts*, 9(4):368.

Zakzanis, K. K., Mcdonald, K., Troyer, A. K., Zakzanis, K. K., Mcdonald, K., Troyer, A. K., Zakzanis, K. K., Mcdonald, K., and Troyer, A. K. (2011). Component analysis of verbal fluency in patients with mild traumatic brain injury Component analysis of verbal fluency in patients with mild traumatic brain injury. *Journal of clinical and experimental neuropsychology*, 3395.

Zarino, B., Crespi, M., Launi, M., and Casarotti, A. (2014). A new standardization of semantic verbal fluency test. *Neurological Sciences*, 35:1405–1411.

Zhao, Q., Guo, Q., and Hong, Z. (2013). Clustering and switching during a semantic verbal fluency test contribute to differential diagnosis of cognitive impairment. *Neuroscience bulletin*, 29(1):75–82.

Zimmer, P., Bloch, W., Schenk, A., Oberste, M., Riedel, S., Kool, J., Langdon, D., Dalgas, U., Kesselring, J., and Bansi, J. (2018). High-intensity interval exercise improves cognitive performance and reduces matrix metalloproteinases-2 serum levels

in persons with multiple sclerosis: A randomized controlled trial. *Multiple Sclerosis Journal*, 24(12):1635–1644.

Zimmermann, N., Parente, M. A. d. M. P., Joanette, Y., and Fonseca, R. P. (2014). Unconstrained, phonemic and semantic verbal fluency: age and education effects, norms and discrepancies. *Psicologia: Reflexão e Crítica*, 27:55–63.

Zipf, G. K. (1936). *The psycho-biology of language: An introduction to dynamic philology*. Routledge.

Appendix A

Supplementary Materials

A.1 Abstract of the Psychonomic Society: 61st Annual Meeting

Abstracts
OF THE PSYCHONOMIC SOCIETY
Volume 25 • November 2020

4:00-6:00 PM (2247)

Turkish Semantic Verbal Fluency Performance in Native Speakers: A Systematic Review. RABIA YASA KOSTAS, SARAH MACPHERSON, and MARIA WOLTERS, *University of Edinburgh* - The semantic verbal fluency test is widely used to examine executive function and semantic abilities. Sematic verbal fluency performance is influenced by many factors, ranging from using a second language to cognitive impairment. Since the test is relatively short, it is a popular candidate for brief semi-automated screening tools. However, fo r clinical use, separate norms should be established for each language. We conducted a systematic review of studies that report verbal fluency performance for native Turkish speakers. Web of Science, Medline, Psycinfo, Embase, and Turkish scholarly databases were searched, in addition to the grey literature. Studies were not limited to specific disorders or demographics. We analysed study designs, scoring methods, and results reported. We only found one study that contained normative data, and the remaining relevant studies were too diverse to establish metanorms. We discuss the implications for deploying verbal fluency tasks in semi-automated cognitive screening tests for Turkish.

Email: Rabia Yasa Kostas,

PSYCHONOMIC SOCIETY®
A PSYCHONOMIC SOCIETY PUBLICATION
www.psychonomic.org

A.2 PROSPERO: International Prospective Register of Systematic Reviews

NIHR | National Institute for Health Research

PROSPERO
International prospective register of systematic reviews

Systematic review of Turkish semantic verbal fluency performance in native speakers

Rabia Yasa Kostas, Maria K. Wolters, Sarah E. MacPherson

Citation

Rabia Yasa Kostas, Maria K. Wolters, Sarah E. MacPherson. Systematic review of Turkish semantic verbal fluency performance in native speakers. PROSPERO 2020 CRD42020201585
Available from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020201585

Review question

To investigate what is known about semantic verbal fluency performance for native Turkish speakers.

PICOS:

P (Population): native Turkish speakers;

I (Intervention): administration of semantic verbal fluency test, regardless of category (animal names, supermarket items, etc.);

C (Comparison): age, gender, and education level, health conditions, bilingualism;

O (Outcome): number of words produced, including any other relevant metrics (e.g. cluster size, number of switches);

S (Study): any psychological or medical study that reports verbal fluency data for native Turkish speakers.

Searches

The literature search will be conducted on the following electronic databases: Web of Science, PsycINFO, EMBASE, and MEDLINE.

Since we are looking for studies regarding to the Turkish population, the following Turkish databases also will be searched:

Turkish Academic Network and Information Centre (ULAKBIM-TR directory), Databases of National Thesis Center of the Council of Higher Education , DergiPark.

The literature search will be done without any restrictions on the year of publication.

The key terms are: semantic verbal fluency, verbal fluency, category fluency, semantic fluency, animal naming, animal fluency, COWAT, Controlled Oral Word Association Test and Turkish language or Turkish population.

Example search strategy for Web of Science:

TS=((semantic fluency OR category fluency OR verbal fluency OR semantic verbal fluency OR COWAT OR Controlled Oral Word Association Test OR animal fluency OR animal naming) AND turk*)

TS= Searching Topic, Title, Abstract, Author Keywords, KeyWords Plus

Types of study to be included

The review will focus on studies that have reported semantic verbal fluency data in native Turkish speakers, both those that report norms and those that include cross-sectional or longitudinal comparisons of semantic verbal fluency performance.

There will be no restrictions on the study design.

There will be no restrictions on the settings of the studies.

Condition or domain being studied

We want to investigate what is known about semantic/category verbal fluency performance in native Turkish speakers.

The semantic verbal fluency test examines executive function and semantic abilities. The duration of the test is usually 60 seconds (sometimes 90 s) and during this short time subjects produce as many words as possible in a given category (animal naming, etc.). Semantic verbal fluency scores are part of many standard cognitive tests, such as the Addenbrooke's Cognitive Examination (Mathuranath et al., 2000) . People with mental health conditions, such as depression, and people with cognitive impairment, often score worse on this test than healthy controls. There has been a surge of interest in developing automatic scoring techniques to help deploy the test for large-scale screening and in telemedicine. Most of the relevant work has been on a small number of languages. Other languages, such as Turkish, have been comparatively neglected. Therefore, the aim is to investigate what is known about semantic verbal fluency performance for native Turkish speakers.

Participants/population

Native mono- or bilingual Turkish speakers.

Intervention(s), exposure(s)

Semantic verbal fluency test administration using any semantic category.

Comparator(s)/control

Psychiatric diagnoses, neurodevelopmental disorders, neurodegenerative diseases, normative data, bilingualism, age.

Main outcome(s)

Semantic verbal fluency test score in people who are native Turkish speakers.

We are interested in (a) variations in performance of healthy speakers (e.g., monolingual versus bilingual; older versus younger) and (b) differences in performance between healthy speakers and speakers with an illness or impairment (e.g., dementia versus control).

* Measures of effect

The standard measurements of semantic verbal fluency performance is number of total words produced. We will also include any other metrics that are reported, e.g., some authors report the number of clusters of related words, the mean size of these clusters, and the number of switches between clusters.

If other metrics have been analysed, these will also be extracted.

Additional outcome(s)

None.

* Measures of effect

Not applicable.

Data extraction (selection and coding)

The search results will be exported as a BibTeX or RIS format, and the review will be managed using Rayyan.

The studies which meet the inclusion criteria will be selected for use in the review.

Studies will be included if they report semantic verbal fluency data for native speakers of Turkish, together with minimal demographic information such as age.

Data will then be extracted from the studies selected for inclusion.

Studies will be coded for design (longitudinal / cross sectional), parameters of semantic fluency tasks administered (duration, semantic categories used), scoring, and, where relevant, participant groups used for reporting (e.g. monolingual versus bilingual, participants with dementia versus participants without dementia).

Risk of bias (quality) assessment

Since we are mainly interested in the verbal fluency results, and these may not have been a main outcome of the study, we will not conduct a formal bias assessment of the studies themselves. Rather, we will note the level of detail with which the data collection and analysis procedure for verbal fluency scores has been reported, and will summarise our findings in an overall assessment of studies.

Studies will receive one point for each relevant detail (e.g., whether words were audio recorded or recorded in writing; whether a reference is given that describes the exact scoring method; etc.).

For each study, two members of the review team will independently perform the quality assessment. Disagreements will be resolved by discussion between all three members of the review team (RYK, MW, SMcP).

Data synthesis will be independent of the quality assessment.

Strategy for data synthesis

We plan a descriptive quantitative synthesis of the Turkish semantic verbal fluency (SVF) data. From each study, we will extract the mean, median, and standard deviation for the following scores:

Number of words produced (overall score);

Number of clusters, number of switches, mean cluster size (if computed).

We will extract SVF scores both overall and for the following subcategories, if they are mentioned in the paper:

Semantic category used (animals, supermarket items, etc.);

Population (e.g. age, gender, education level);

Comparators used in analysis (e.g., monolingual versus bilingual; healthy controls versus people with Alzheimer's disease).

We will also summarise how often other scores were given, such as number of words produced in the first 15 seconds, and time between words, but we will not use descriptive statistics.

Finally, we will produce a table that summarises the level of detail with which data collection and analysis methods were reported.

Where a paper reports differences between groups on SVF performance, we will tabulate the group comparisons considered, the statistical tests used, and whether differences were significant or not.

Since we are including all studies of SVF that report Turkish data, it does not make sense to perform meta-analyses for each of the group comparisons, or to attempt to establish a definitive range of normal performance.

We expect there to be such substantial variation in study goals and study comparators that fitting a formal statistical model will not be indicated. Instead, we aim to highlight gaps in the data, and compare the existing literature for Turkish to the literature for languages such as English and German, where meta-analyses are a lot more feasible.

To conduct the review, we will use Rayyan QCRI.

All statistical analyses will be performed using R (tidyverse packages) by two reviewers independently.

Differences will be reconciled by all three reviewers.

Analysis of subgroups or subsets
None planned.

Contact details for further information
Rabia Yasa Kostas

Organisational affiliation of the review
School of Informatics, The University of Edinburgh

https://www.ed.ac.uk/informatics/

Review team members and their organisational affiliations
Mrs Rabia Yasa Kostas. School of Informatics, The University of Edinburgh
Dr Maria K. Wolters. School of Informatics, The University of Edinburgh
Dr Sarah E. MacPherson. School of Philosophy, Psychology and Language Sciences, The University of Edinburgh

Type and method of review
Diagnostic, Systematic review

Anticipated or actual start date
01 September 2020

Anticipated completion date
28 February 2021

Funding sources/sponsors
Graduate Scholarship: Republic of Turkey Ministry of National Education

Conflicts of interest

Language
English, Turkish

Country
Scotland

Stage of review
Review Ongoing

Subject index terms status
Subject indexing assigned by CRD

Subject index terms

Diagnostic Techniques, Neurological; Humans; Language; Language Tests; Mental Disorders; Mental Health; Neurologic Manifestations; Public Health; Semantics; Speech Disorders; Turkey; Verbal Behavior

Date of registration in PROSPERO

27 October 2020

Date of first submission

05 August 2020

Stage of review at time of this submission

Stage	Started	Completed
Preliminary searches	Yes	No
Piloting of the study selection process	No	No
Formal screening of search results against eligibility criteria	No	No
Data extraction	No	No
Risk of bias (quality) assessment	No	No
Data analysis	No	No

The record owner confirms that the information they have supplied for this submission is accurate and complete and they understand that deliberate provision of inaccurate information or omission of data may be construed as scientific misconduct.

The record owner confirms that they will update the status of the review when it is completed and will add publication details in due course.

Versions

27 October 2020

A.3 ETHICS

PPLS RESEARCH ETHICS COMMITTEE
SCHOOL of PHILOSOPHY, PSYCHOLOGY and LANGUAGE SCIENCES
The University of Edinburgh
Dugald Stewart Building
3 Charles Street
Edinburgh EH8 9AD
Telephone +44 (0) 131 651 1761
Email PPLS.EthicsAdmin@ed.ac.uk

9 September 2021

Ethics proposal 383-2021/2, entitled Turkish Semantic Verbal Fluency: Online Data Collection and Computational Analysis and submitted by Rabia Yasa Kostas, Rabia Yasa Kostas, Danyi He, Dr Sarah E MacPherson and Maria Wolters has been approved by the PPLS Research Ethics Committee per the Department's ethics regulations.

The following files were uploaded with the application:

Filename: Final_webexperiment_Identifiable_consent_gdpr.docx
Date: 22 Aug 2021 05:28 PM
Purpose: Information Sheet
Note: Information sheet and consent form

Filename: Final_webexperiment_Identifiable_consent_gdpr.docx
Date: 22 Aug 2021 05:28 PM
Purpose: Consent Sheet
Note: Information sheet and consent form

Filename: ReviewComments-383-2021_1.pdf
Date: 31 Aug 2021 02:42 PM
Purpose: (Reply to)/PPLSREC Review
Note: Comments from reviewer in response to your submission.

A.4 Participant Information Sheet (PIS form)

THE UNIVERSITY of EDINBURGH
School of Philosophy, Psychology and Language Sciences

Information sheet for participants

Study title:	Collection of semantic verbal fluency performance in native Turkish speakers
Investigators:	Dr. Maria Wolters (Principal Investigator), Dr Sarah MacPherson (co-investigator), Rabia Yasa Kostas (co-investigator)
Researcher collecting data:	Rabia Yasa Kostas

What is this document? This document explains what kind of study we're doing, what your rights are, and what will be done with your data. You should print this page for your records.

Nature of the study. You are invited to participate in a study which involves collecting a large set of data on semantic verbal fluency in Turkish. Semantic Verbal Fluency is a short task that is part of many psychological studies and involves naming as many words from a given category as possible in 60 seconds. In this study, we will ask you to complete the Semantic Verbal Fluency Task for three categories, including animals. Your responses for each category will be audio recorded, and the recordings will be transcribed. We will also ask you some questions about your background, such as age, gender, or language background. Your session should last for up to 15 minutes (10 for the questionnaire, 5 for the three Semantic Verbal Fluency tasks). You will be given full instructions shortly.

Compensation. There is no compensation for your participation in this study.

Risks and benefits. There are no known risks to participation in this study. There are no tangible benefits to you, however you will be contributing to our knowledge about semantic verbal fluency in Turkish, which will benefit everyone doing psychological research with Turkish people.

Confidentiality and use of data. All the information we collect during the course of the research will be processed in accordance with Data Protection Law. In order to safeguard your privacy, we will never share personal information (like names or dates of birth) with anyone outside the research team. Your data will be referred to by a unique participant number rather than by name. We will use this unique number to link the information about your background with the tool that collects semantic verbal fluency data. We will store any personal data, such as audio recordings, using the University of Edinburgh DataShare service. The anonymised data collected during this study will be used for research purposes. With your permission, your anonymized, transcribed data can be shared with other researchers. We will not share your audio recordings.

What are my data protection rights? The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure, and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments, and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Voluntary participation and right to withdraw. Your participation is voluntary, and you may withdraw from the study at any time and for any reason. If you withdraw from the study during or after data gathering, we will delete your data and there is no penalty or loss of benefits to which you are otherwise entitled.

If you have any questions about what you've just read, please feel free to ask, or contact us later. You can contact us by email at . This project has been approved by PPLS Ethics committee. If you have questions or comments regarding your rights as a participant, they can be contacted at 0131 650 4020 or ppls.ethics@ed.ac.uk.

By responding "yes", you consent to the following:

1. **I agree to participate in this study.**
2. I confirm that I have read and understood **how my data will be stored and used**.
3. I understand that I have the **right to terminate this session** at any point
4. I understand that my anonymized data may be shared with other researchers.

A.5 Spreadsheet of number of salient annotation elements in Turkish dataset

ID	Category	Number of Perseverations	Number of Category Violations	Number of Helper speaking	Number of Phrases-Extra speech	ID	Category	Number of Perseverations	Number of Category Violations	Number of Helper speaking	Number of Phrases-Extra speech
110726	Animal	0	0	0	0	557453	Animal	0	0	0	0
113362	Animal	1	0	0	0	579558	Animal	0	0	0	0
129534	Animal	11	0	0	0	587161	Animal	0	0	0	0
134738	Animal	1	0	0	0	596476	Animal	1	0	0	0
141269	Animal	2	0	0	0	605451	Animal	1	0	0	0
158057	Animal	0	0	0	0	610796	Animal	3	0	0	0
167708	Animal	2	0	0	0	615759	Animal	0	0	0	0
171901	Animal	1	0	0	0	634211	Animal	3	0	0	1
172224	Animal	0	0	0	0	641102	Animal	2	0	0	0
188744	Animal	0	0	0	0	656805	Animal	4	0	0	0
207259	Animal	0	0	0	1	661525	Animal	1	0	0	0
208272	Animal	5	0	0	0	691042	Animal	1	0	0	0
208452	Animal	1	0	0	0	704626	Animal	1	0	0	0
232576	Animal	0	0	0	0	712233	Animal	0	0	0	2
234153	Animal	0	0	0	1	712635	Animal	3	0	0	0
240688	Animal	0	0	6	0	714942	Animal	1	0	0	0
241654	Animal	0	0	0	0	715519	Animal	0	0	0	0
258508	Animal	0	0	0	0	728594	Animal	0	0	0	1
266469	Animal	4	0	0	0	731118	Animal	0	0	0	0
269103	Animal	1	0	0	0	732954	Animal	0	0	0	0
279304	Animal	2	0	0	0	743820	Animal	0	0	0	0
281672	Animal	0	0	0	0	747878	Animal	0	0	0	0
289868	Animal	1	0	0	0	758394	Animal	1	0	0	1
314479	Animal	0	0	0	0	762198	Animal	0	0	0	0
314880	Animal	0	0	0	0	763105	Animal	0	0	0	0
319968	Animal	1	0	0	0	769088	Animal	2	0	0	1
323439	Animal	1	0	0	1	770940	Animal	6	0	0	2
335556	Animal	0	0	0	2	781445	Animal	0	0	0	3
337732	Animal	0	0	0	0	782771	Animal	1	0	0	0
343315	Animal	0	0	0	0	841585	Animal	1	0	0	0
355457	Animal	1	1	0	0	844303	Animal	0	0	0	0
355521	Animal	1	0	0	0	863146	Animal	0	0	0	0
358075	Animal	1	0	0	0	876355	Animal	0	0	1	0
361712	Animal	0	0	0	0	917926	Animal	1	0	0	4
385025	Animal	0	0	0	3	940686	Animal	1	0	0	0
390449	Animal	1	0	0	1	954491	Animal	3	0	3	0
400826	Animal	0	0	0	0	959630	Animal	1	0	0	0
402926	Animal	0	0	0	0	960844	Animal	4	0	0	1
453717	Animal	0	0	0	0	967278	Animal	1	0	0	0
493359	Animal	3	0	0	0	977796	Animal	0	0	0	1
497136	Animal	0	0	0	0	980200	Animal	0	0	0	1
500922	Animal	2	0	0	1	990429	Animal	6	0	0	0
501061	Animal	0	0	0	2	153787	Animal	0	0	0	0
501255	Animal	0	0	0	1	198916	Animal	0	3	0	0
501772	Animal	0	0	0	1	233592	Animal	2	0	5	0
509719	Animal	5	0	0	0	306870	Animal	0	0	0	0
516419	Animal	0	0	0	0	460944	Animal	2	0	0	0
528497	Animal	2	0	0	0	582397	Animal	4	0	0	0
534368	Animal	0	0	0	3	583972	Animal	4	0	0	0
555225	Animal	0	0	0	0	587309	Animal	0	0	0	0
555902	Animal	0	0	0	1	776693	Animal	0	0	0	2
819423	Animal	10	0	0	0	781673	Animal	6	0	0	0
826835	Animal	1	0	0	0						

ID	Category	Number of Perseverations	Number of Category Violations	Number of Helper speaking	Number of Phrases-Extra speech	ID	Category	Number of Perseverations	Number of Category Violations	Number of Helper speaking	Number of Phrases-Extra speech
135787	Vegetables and Fruits	8	0	0	2	500922	Vegetables and Fruits	1	0	0	0
403817	Vegetables and Fruits	0	0	0	2	501061	Vegetables and Fruits	0	0	0	1
439896	Vegetables and Fruits	0	0	0	0	501255	Vegetables and Fruits	3	0	0	0
521845	Vegetables and Fruits	0	0	0	0	501772	Vegetables and Fruits	0	0	0	1
563838	Vegetables and Fruits	0	0	0	0	509719	Vegetables and Fruits	2	0	0	0
583972	Vegetables and Fruits	1	0	0	0	516419	Vegetables and Fruits	0	0	0	0
587309	Vegetables and Fruits	1	0	0	2	528497	Vegetables and Fruits	5	0	0	0
643578	Vegetables and Fruits	9	0	0	0	534368	Vegetables and Fruits	0	1	0	0
713964	Vegetables and Fruits	1	0	0	6	555225	Vegetables and Fruits	4	0	0	1
717733	Vegetables and Fruits	9	0	0	2	555902	Vegetables and Fruits	2	0	0	0
763804	Vegetables and Fruits	0	0	0	0	557453	Vegetables and Fruits	1	0	0	0
819423	Vegetables and Fruits	2	0	0	0	579558	Vegetables and Fruits	0	0	0	0
110726	Vegetables and Fruits	2	0	0	0	587161	Vegetables and Fruits	1	1	0	0
113362	Vegetables and Fruits	3	0	0	0	596476	Vegetables and Fruits	0	0	0	0
129534	Vegetables and Fruits	3	0	0	0	605451	Vegetables and Fruits	1	0	0	0
134738	Vegetables and Fruits	1	0	0	0	610796	Vegetables and Fruits	3	0	0	0
141269	Vegetables and Fruits	2	0	0	0	615759	Vegetables and Fruits	2	0	0	1
158057	Vegetables and Fruits	4	0	0	0	634211	Vegetables and Fruits	4	0	0	1
167708	Vegetables and Fruits	1	0	0	0	641102	Vegetables and Fruits	0	0	0	0
171901	Vegetables and Fruits	0	0	0	0	656805	Vegetables and Fruits	3	0	0	0
172224	Vegetables and Fruits	1	0	0	1	661525	Vegetables and Fruits	1	0	0	0
188744	Vegetables and Fruits	0	0	0	0	691042	Vegetables and Fruits	0	0	0	0
207259	Vegetables and Fruits	0	0	0	0	704626	Vegetables and Fruits	1	0	0	0
208272	Vegetables and Fruits	0	0	0	0	712233	Vegetables and Fruits	0	0	0	1
208452	Vegetables and Fruits	0	0	0	0	712635	Vegetables and Fruits	4	0	0	0
232576	Vegetables and Fruits	1	0	0	0	714942	Vegetables and Fruits	1	0	0	0
234153	Vegetables and Fruits	0	0	0	0	715519	Vegetables and Fruits	1	0	0	0
240688	Vegetables and Fruits	0	0	4	0	728594	Vegetables and Fruits	0	0	0	0
241654	Vegetables and Fruits	2	0	0	0	731118	Vegetables and Fruits	1	0	0	0
258508	Vegetables and Fruits	0	0	0	0	732954	Vegetables and Fruits	0	0	0	0
266469	Vegetables and Fruits	2	0	0	0	743820	Vegetables and Fruits	1	0	0	0
269103	Vegetables and Fruits	0	0	0	0	747878	Vegetables and Fruits	1	0	0	0
279304	Vegetables and Fruits	3	0	0	0	758394	Vegetables and Fruits	0	0	0	2
281672	Vegetables and Fruits	0	0	0	0	762198	Vegetables and Fruits	1	0	0	1
289868	Vegetables and Fruits	4	0	0	0	763105	Vegetables and Fruits	0	0	0	0
314479	Vegetables and Fruits	0	0	0	0	769088	Vegetables and Fruits	2	0	0	0
314880	Vegetables and Fruits	0	0	0	0	770940	Vegetables and Fruits	2	0	0	1
319968	Vegetables and Fruits	2	1	0	0	781445	Vegetables and Fruits	2	0	0	0
323439	Vegetables and Fruits	5	0	0	0	782771	Vegetables and Fruits	3	0	0	0
335556	Vegetables and Fruits	2	0	0	0	841585	Vegetables and Fruits	0	0	0	0
337732	Vegetables and Fruits	0	0	0	0	844303	Vegetables and Fruits	1	0	0	1
343315	Vegetables and Fruits	1	0	0	0	863146	Vegetables and Fruits	1	0	0	0
355457	Vegetables and Fruits	1	0	0	0	876355	Vegetables and Fruits	0	0	1	0
355521	Vegetables and Fruits	1	0	0	0	917926	Vegetables and Fruits	1	0	0	3
358075	Vegetables and Fruits	4	0	0	0	940686	Vegetables and Fruits	4	0	0	2
361712	Vegetables and Fruits	3	0	0	0	954491	Vegetables and Fruits	1	0	4	0
385025	Vegetables and Fruits	1	0	0	3	959630	Vegetables and Fruits	5	0	0	0
390449	Vegetables and Fruits	3	0	0	0	960844	Vegetables and Fruits	0	0	0	1
400826	Vegetables and Fruits	1	0	0	0	967278	Vegetables and Fruits	0	0	0	0
402926	Vegetables and Fruits	1	0	0	0	977796	Vegetables and Fruits	3	0	0	0
453717	Vegetables and Fruits	1	0	0	0	980200	Vegetables and Fruits	1	0	0	0
493359	Vegetables and Fruits	7	3	0	0	990429	Vegetables and Fruits	2	0	0	0
497136	Vegetables and Fruits	14	0	0	0						

ID	Category	Number of Perseverations	Number of Category Violations	Number of Helper speaking	Number of Phrases-Extra speech
294937	Supermarket Items	1	0	0	0
403817	Supermarket Items	0	0	0	2
439896	Supermarket Items	0	0	0	0
521845	Supermarket Items	0	0	0	0
563838	Supermarket Items	5	0	0	0
643578	Supermarket Items	0	0	0	0
713964	Supermarket Items	1	0	0	6
763804	Supermarket Items	0	0	0	0
110726	Supermarket Items	0	0	0	0
113362	Supermarket Items	2	0	0	0
129534	Supermarket Items	6	0	0	0
134738	Supermarket Items	2	0	0	0
141269	Supermarket Items	0	0	0	0
158057	Supermarket Items	2	0	0	0
167708	Supermarket Items	0	0	0	0
171901	Supermarket Items	2	0	0	0
172224	Supermarket Items	1	0	0	0
188744	Supermarket Items	0	0	0	0
207259	Supermarket Items	0	0	0	1
208272	Supermarket Items	1	0	0	0
208452	Supermarket Items	2	0	0	0
232576	Supermarket Items	4	0	0	0
234153	Supermarket Items	0	0	0	0
240688	Supermarket Items	0	0	2	0
241654	Supermarket Items	1	0	0	0
258508	Supermarket Items	0	0	0	0
266469	Supermarket Items	2	0	0	0
269103	Supermarket Items	0	0	0	0
279304	Supermarket Items	0	0	0	0
281672	Supermarket Items	1	0	0	0
289868	Supermarket Items	4	0	0	0
314479	Supermarket Items	0	0	0	2
314880	Supermarket Items	0	0	0	0
319968	Supermarket Items	0	0	0	0
323439	Supermarket Items	0	0	0	0
335556	Supermarket Items	0	0	0	0
337732	Supermarket Items	0	0	0	0
343315	Supermarket Items	0	0	0	0
355457	Supermarket Items	5	0	0	0
355521	Supermarket Items	4	0	0	0
358075	Supermarket Items	2	0	0	0
361712	Supermarket Items	3	0	0	0
385025	Supermarket Items	0	0	0	2
390449	Supermarket Items	1	0	0	0
400826	Supermarket Items	0	0	0	0
402926	Supermarket Items	0	0	0	0
453717	Supermarket Items	0	0	0	0
493359	Supermarket Items	4	0	0	0
497136	Supermarket Items	6	0	0	0
500922	Supermarket Items	0	0	0	0
501061	Supermarket Items	2	0	0	4
501255	Supermarket Items	0	0	0	0
501772	Supermarket Items	1	0	0	0
509719	Supermarket Items	0	0	0	0
516419	Supermarket Items	0	0	0	0
528497	Supermarket Items	0	0	0	0
534368	Supermarket Items	0	0	0	1
555225	Supermarket Items	1	0	0	1
555902	Supermarket Items	1	0	0	0
557453	Supermarket Items	0	0	0	0
579558	Supermarket Items	1	0	0	0
587161	Supermarket Items	0	0	0	0
596476	Supermarket Items	1	0	0	0
605451	Supermarket Items	1	0	0	0
610796	Supermarket Items	0	0	0	0
615759	Supermarket Items	1	0	0	1
634211	Supermarket Items	2	0	0	1
641102	Supermarket Items	0	0	0	0
656805	Supermarket Items	0	0	0	0
661525	Supermarket Items	1	0	0	0
691042	Supermarket Items	0	0	0	0
704626	Supermarket Items	0	0	0	0
712233	Supermarket Items	0	0	0	0
712635	Supermarket Items	0	0	0	0
714942	Supermarket Items	1	0	0	0
715519	Supermarket Items	0	0	0	0
728594	Supermarket Items	1	0	0	0
731118	Supermarket Items	3	0	0	1
732954	Supermarket Items	0	0	0	0
743820	Supermarket Items	1	0	0	0
747878	Supermarket Items	0	0	0	0
758394	Supermarket Items	1	0	0	2
762198	Supermarket Items	0	0	0	0
763105	Supermarket Items	0	0	0	0
769088	Supermarket Items	1	0	0	0
770940	Supermarket Items	1	0	0	2
781445	Supermarket Items	0	0	0	0
782771	Supermarket Items	0	0	2	0
841585	Supermarket Items	0	0	0	0
844303	Supermarket Items	2	0	0	0
863146	Supermarket Items	1	0	0	0
876355	Supermarket Items	1	0	1	0
917926	Supermarket Items	0	0	0	2
940686	Supermarket Items	2	0	0	0
954491	Supermarket Items	0	0	1	0
959630	Supermarket Items	0	0	0	0
960844	Supermarket Items	0	0	0	0
967278	Supermarket Items	1	0	0	0
977796	Supermarket Items	2	0	0	1
980200	Supermarket Items	0	0	0	0
990429	Supermarket Items	0	0	0	0

A.6 Original Troyer Taxonomy

Main groups	Sub groups	Sub group animals
Living Environment	Africa	aardvark, antelope, buffalo, camel, chameleon, cheetah, chimpanzee, cobra, eland,elephant, gazelle, giraffe, gnu, gorilla, hippopotamus, hyena, impala, jackal, lemur, leopard,lion, manatee, mongoose, monkey, ostrich, panther, rhinoceros, tiger, wildebeest, warthog,zebra
Living Environment	Australia	emu, kangaroo, kiwi, opossum, platypus, Tasmanian devil, wallaby, wombat
Living Environment	Arctic/Far North	auk, caribou, musk ox, penguin, polar bear, reindeer, seal
Living Environment	Farm	chicken, cow, donkey, ferret, goat, horse, mule, pig, sheep, turkey
Living Environment	North America	badger, bear, beaver, bobcat, caribou, chipmunk, cougar, deer, elk, fox,moose, mountain lion, puma, rabbit, raccoon, skunk, squirrel, wolf
Living Environment	Water	alligator, auk, beaver, crocodile, dolphin, fish, frog, lobster, manatee, muskrat, newt,octopus, otter, oyster, penguin, platypus, salamander, sea lion, seal, shark, toad, turtle, whale
Human Use	Beasts of burden	camel, donkey, horse, llama, ox
Human Use	Fur	beaver, chinchilla, fox, mink, rabbit
Human Use	Pets	budgie, canary, cat, dog, gerbil, golden retriever, guinea pig, hamster, parrot, rabbit
Zoological Categories	Bird	budgie, condor, eagle, finch, kiwi, macaw, parrot, parakeet, pelican, penguin, robin,toucan, woodpecker
Zoological Categories	Bovine	bison, buffalo, cow, musk ox, yak
Zoological Categories	Canine	coyote, dog, fox, hyena, jackal, wolf
Zoological Categories	Deer	antelope, caribou, eland, elk, gazelle, gnu, impala, moose, reindeer, wildebeest
Zoological Categories	Feline	bobcat, cat, cheetah, cougar, jaguar, leopard, lion, lynx, mountain lion, ocelot,panther, puma, tiger
Zoological Categories	Fish	bass, guppy, salmon, trout
Zoological Categories	Insect	ant, beetle, cockroach, flea, fly, praying mantis
Zoological Categories	Insectivores	aardvark, anteater, hedgehog, mole, shrew
Zoological Categories	Primate	ape, baboon, chimpanzee, gibbon, gorilla, human, lemur, marmoset, monkey,orangutan, shrew
Zoological Categories	Rabbit	Coney, hare, pika, rabbit
Zoological Categories	Reptile / Amphibian	alligator, chameleon, crocodile, frog, gecko, iguana, lizard, newt,salamander, snake, toad, tortoise, turtle
Zoological Categories	Rodent	beaver, chinchilla, chipmunk, gerbil, gopher, groundhog, guinea pig, hamster,hedgehog, marmot, mole, mouse, muskrat, porcupine, rat, squirrel, woodchuck
Zoological Categories	Weasel	badger, ferret, marten, mink, mongoose, otter, polecat, skunk

Table A.1: Original Troyer taxonomy with animal names. The list gathered from the article published by Troyer et al. (1997), which they proposed Clustering ans Switching components and scoring rules.

A.7 Spanish version of Troyer Taxonomy

Spanish Expanded Animal Taxonomy

Bioma
África: oso hormiguero, antílope, búfalo, camello, camaleón, guepardo, chimpancé, cobra, tierra, elefante, gacela, jirafa, ñu, gorila, hipopótamo, hiena, impala, chacal, lémur, leopardo, león, manatí, mangosta, mono, avestruz, pantera, rinoceronte, tigre, ñu, jabalí, cebra, **Mico**
Australia: emú, canguro, kiwi, zarigüeya, ornitorrinco, demonio de Tasmania, wallaby, wómbat
Ártico / Norte lejano: alca, caribú, buey almizclero, pingüino, oso polar, reno, foca
Granja: pollo, vaca, burro, hurón, cabra, caballo, mula, cerdo, oveja, pavo, **gallina, ganso, pato, pisco, pollito, asno, chivo, Yegua, Toro, Marrano, gallo, becerro, Ternero**
América del Norte: tejón, oso, castor, lince, caribú, ardilla listada, puma, ciervo, alce, zorro, alce, león de montaña, puma, conejo, mapache, mofeta, ardilla, lobo, **armadillo**
Agua: caimán, auk, castor, cocodrilo, delfín, pez, rana, langosta, manatí, rata almizclera, tritón, pulpo, nutria, ostra, pingüino, ornitorrinco, salamandra, león marino, foca, tiburón, sapo, tortuga, ballena, , **caiman, cangrejo, caballo_de_mar, estrella_de_mar**

Uso humano
Bestias de carga: camello, burro, caballo, llama, buey, **asno, Yegua, Toro**
Pelaje: castor, chinchilla, zorro, visón, conejo
Mascotas: periquito, canario, gato, perro, jerbo, golden retriever, conejillo de indias, hámster, loro, conejo, **curi**

Zoologica
Pájaro: periquito, cóndor, águila, pinzón, kiwi, guacamayo, loro, perico, pelícano, pingüino, petirrojo, tucán, pájaro carpintero, **ave, gorrión, pájaro, paloma, colibrí, búho, buitre, cigüeña, codorniz, Águila, Sinsonte, Tórtola, Turpial, Mirla, Lechuza, Halcón, guacamaya, golondrina, gavilán, gallinazo, condor, murciélago, cisne,**
Bovinos: bisonte, búfalo, vacas, bueye almizclero, yak, **becerro, Ternero**
Canino: coyote, perro, zorro, hiena, chacal, lobo
Ciervo: antílope, caribú, eland, alce, gacela, ñu, impala, alce, reno, ñu
Felino: gato montés, gato, guepardo, puma, jaguar, leopardo, león, lince, león de montaña, ocelote, pantera, puma, tigre
Pescado: lubina, guppy, salmón, trucha, **Pescado**
Insecto: hormiga, escarabajo, cucaracha, pulga, mosca, mantis religiosa, **piojo, abeja, avispa, chinche, Cucarrón, Libélula, Zancudo, mosquito, Mariquita, Mariposa, Luciérnaga, grillo, garrapata, escorpión, gusano, lombriz, alacrán, araña, ciempiés,**
Insectivos: oso hormiguero, oso hormiguero, erizo, topo, musaraña
Primado: mono, babuino, chimpancé, gibón, gorila, humano, lémur, tití, mono, orangután, musaraña, **Mico**
Conejo: conejo, liebre, pika, conejo
Reptiles / anfibios: caimán, camaleón, cocodrilo, rana, geco, iguana, lagarto, tritón, salamandra, serpiente, sapo, tortuga, tortuga, **culebra, caiman, Lagartijo, Lagartija,**
Roedor: castor, chinchilla, ardilla listada, jerbo, gopher, marmota, conejillo de indias, hámster, erizo, marmota, mole, ratón, rata almizclera, puercoespín, rata, ardilla, marmota, **ratón, chucha, Guagua, curi**
Comadreja: tejón, hurón, marta, visón, mangosta, nutria, turón, mofeta
imaginación:beast
dinosaurio:dinosaurio
mamut:mamut

Note: Animal names written in blue are not included in the original Troyer taxonomy. The list is an expanded version of the original Troyer taxonomy with the animals in the data set that we used in Spanish study Chapter-5.

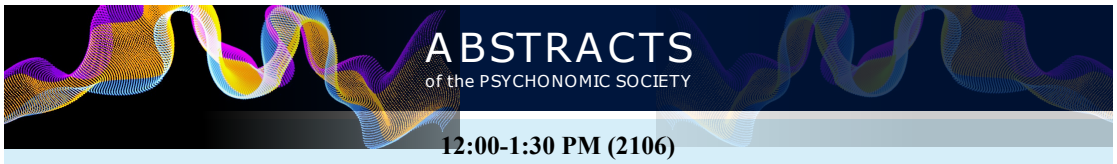
A.8 Full table of Descriptive statistic of Word2Vec models of Colombian-Spanish SVF analysis study

					Snowball Stem								Patternlib Lemma								Spacy Lemma							
Model		**hyperparameters**			**Number of switches**				**Mean Cluster size**				**Number of switches**				**Mean Cluster size**				**Number of switches**				**Mean Cluster size**			
	Th	***d***	**w**	**f**	**Mean**	**Mdn**	**Max**	**Min**	**Mean**	**Mdn**	**Max**	**Min**	**Mean**	**Mdn**	**Max**	**Min**	**Mean**	**Mdn**	**Max**	**Min**	**Mean**	**Mdn**	**Max**	**Min**	**Mean**	**Mdn**	**Max**	**Min**
Word2vec	**0.75**	1000	10	cbow	13.3	14	23	2	1.45	1.42	2.2	1	13.44	14	22	3	1.43	1.4	2.2	1.07	12.86	13	23	2	1.49	1.44	2	1.08
		1000	**10**	**skipgram**	12.95	13	23	2	1.5	1.42	2.3	1.1	13.02	13	24	2	1.5	1.44	2.4	1.08	12.86	13	23	2	1.51	1.44	2.29	1.11
		1000	4	cbow	13.64	14	24	3	1.41	1.37	2.2	1	13.22	13	22	3	1.46	1.39	2.2	1	13.06	13.5	24	2	1.47	1.39	2	1.08
		1000	4	skipgram	12.88	13	24	3	1.5	1.46	2.7	1.1	12.84	12.5	24	2	1.51	1.44	2.25	1.07	12.75	13	22	2	1.52	1.46	2.29	1.11
		600	10	cbow	13.23	14	22	2	1.46	1.41	2.3	1	13.19	13.5	22	2	1.46	1.42	2	1.07	13.12	13	22	2	1.48	1.45	2	1
		600	10	skipgram	13.23	13	23	2	1.46	1.4	2.2	1.1	13.02	13	24	2	1.51	1.42	2.4	1.08	12.81	13	23	2	1.53	1.44	2.67	1.11
		600	4	cbow	13.17	13	22	3	1.45	1.42	2.2	1.1	13.03	13	22	3	1.48	1.42	2.2	1.07	12.75	12.5	23	2	1.52	1.44	2.1	1.08
		600	4	skipgram	12.92	13.5	24	3	1.5	1.44	2.3	1.1	12.7	13	23	2	1.54	1.46	2.29	1.07	12.61	12.5	22	2	1.55	1.44	2.29	1.11
		300	10	cbow	12.83	13	22	2	1.51	1.46	2.2	1.1	13.03	13	22	1	1.5	1.44	3	1.07	12.98	13	22	2	1.5	1.43	2	1
		300	10	skipgram	13.17	14	23	2	1.47	1.4	2.1	1.1	12.97	13	26	2	1.51	1.43	2.4	1.11	12.95	13	25	2	1.51	1.39	2.67	1.11
		300	4	cbow	13.2	13	22	2	1.47	1.41	2.3	1	13.02	13	22	2	1.49	1.41	2.2	1	13.11	13	22	2	1.48	1.42	2.25	1
		300	4	skipgram	12.94	14	23	2	1.52	1.4	2.4	1.1	13.02	13	24	2	1.5	1.42	2.4	1.07	12.7	13	23	2	1.54	1.44	2.4	1.11
	0.50	1000	10	cbow	8.31	8	15	2	2.28	2.11	4	1.4	8.3	8	17	1	2.34	2.17	4.75	1.23	8.31	8	17	1	2.37	2.24	7.5	1.25
		1000	10	skipgram	8.2	8	17	2	2.35	2.18	4.5	1.4	8.19	8	19	2	2.32	2.21	3.67	1.55	8.34	8.5	16	1	2.4	2.04	7.5	1.43
		1000	4	cbow	8.38	8	15	2	2.32	2.1	4.8	1.4	8.41	8	18	1	2.33	2.21	4.75	1.45	8.38	8.5	17	2	2.3	2.16	4.5	1.11
		1000	4	skipgram	8.25	8	16	2	2.28	2.17	4.5	1.4	8.16	8	16	2	2.35	2.22	4.6	1.56	8.22	8	16	2	2.33	2.1	4.5	1.4
		600	10	cbow	8.31	8.5	15	1	2.32	2.17	4.5	1.4	8.59	8	18	1	2.28	2.05	4	1.29	8.31	8	17	1	2.39	2.25	5	1.38
		600	10	skipgram	8.22	8	17	2	2.31	2.15	4.5	1.6	8.11	8	18	2	2.39	2.17	4.75	1.45	8.3	8	16	2	2.37	2.09	5	1.43
		600	4	cbow	8.41	8	15	2	2.3	2.13	4.6	1.4	8.27	8	17	1	2.4	2.19	4.75	1.38	8.27	8	15	1	2.4	2.24	6	1.11
		600	4	skipgram	8.22	8	18	2	2.34	2.2	4.5	1.4	8.17	8	20	1	2.36	2.28	4	1.33	8.23	8	16	1	2.39	2.17	5.5	1.4
		300	10	cbow	8.28	9	14	1	2.32	2.2	3.8	1.4	8.59	8.5	18	1	2.29	2.09	4	1.23	8.38	8	18	1	2.34	2.22	5	1.11
		300	10	skipgram	8.19	8	17	2	2.34	2.17	4.5	1.3	8.16	8	19	1	2.39	2.19	4	1.45	8.23	8	17	2	2.4	2.14	6	1.4
		300	4	cbow	8.38	8	15	1	2.29	2.17	3.8	1.3	8.59	9	17	1	2.29	2.09	4	1.33	8.31	8.5	16	1	2.39	2.3	5	1.11
		300	4	skipgram	8.2	8	16	2	2.37	2.19	4.5	1.3	8.28	8.5	18	1	2.38	2.18	6	1.33	8.27	8	18	2	2.37	2.2	5	1.38
	0.25	1000	10	cbow	4.12	4	8	0	4.48	4.27	11	2.1	3.97	4	9	0	4.99	4.04	12	2	4.09	4	10	0	4.78	4.17	11	2.09
		1000	10	skipgram	4.02	3	10	1	4.71	4.22	11	2	3.94	3.5	8	0	5.11	4	19	2.43	4.02	3.5	11	0	5	4.2	19	2.22
		1000	4	cbow	4.11	4	9	1	4.5	3.93	9.5	1.9	4	4	9	0	4.8	3.93	12	2	4.16	4	10	0	4.89	4.27	15	2
		1000	4	skipgram	3.98	3.5	9	1	4.61	4.36	11	2	4.05	4	9	0	4.64	4	12	2.12	4.05	4	10	0	4.87	4	15	2
		600	10	cbow	4.08	4	8	0	4.47	4.29	11	2.1	4.02	4	10	0	4.73	4.2	11.5	1.82	4.02	4	9	0	5.01	4.22	15	2.12
		600	10	skipgram	4.05	4	9	1	4.71	4	11	1.9	3.94	4	10	0	5.23	4	21	2.27	4.05	4	9	0	4.98	4.17	19	2.22
		600	4	cbow	4.08	4	9	0	4.59	4.2	12	2	4.03	4	11	0	4.72	4.1	12	2	4.09	4	11	0	4.93	4.22	12	1.8
		600	4	skipgram	4.05	4	9	0	4.6	4.25	11	2	3.95	3	10	0	4.97	4.22	12	2.22	4.08	4	10	0	5	4.22	19	2
		300	10	cbow	4.09	4	9	0	4.62	4.1	12	2.1	4.03	4	9	0	4.86	4.22	19	2	4.12	4	11	0	4.85	3.94	12	2.25
		300	10	skipgram	4.02	3.5	9	0	4.96	4.29	19	1.9	3.89	4	9	0	5.27	4.29	21	2.22	4.06	4	9	0	4.98	4.2	19	2.22
		300	4	cbow	4.05	4	9	0	4.65	3.94	12	2.1	4.03	4	10	0	4.8	3.98	12	2	4.17	4	10	0	4.74	4.2	15	1.82
		300	4	skipgram	4.09	3	9	0	4.67	4.22	19	2	3.95	4	11	0	4.85	4.22	16	2.3	4.03	4	9	0	5.12	4	19	2
Bigram					12.8	13	23	2	1.54	1.5	2.6	1.1																
Troyer					12.22	12	22	2	1.57	1.54	3	1.1																

Table A.2: All descriptive statistics (Mean, Median(Mdn), Maximum(Max), Minimum(Min)) from 36 models of Word2Vec with all morphological analyzers:Snowball Stemmer, Patternlib Lemmatizer, Spacy Lemmatizer. Best performing model are bolded.

A.9 Abstract of the Psychonomic Society: 64th Annual Meeting

Abstracts
OF THE PSYCHONOMIC SOCIETY
Volume 25 • November 2023 64th Annual Meeting



12:00-1:30 PM (2106)

Online Collection of Turkish Semantic Verbal Fluency Data: Benefits and Challenges. MARIA K. WOLTERS, *University of Edinburgh & OFFIS* , RABIA YASA KOSTAS, *University of Edinburgh*, DANYI HE, *Honor Device Co*, SARAH E. MACPHERSON, *University of Edinburgh* - We report on a unique data set of Turkish Semantic Verbal Fluency (SVF) data, collected online using the categories animals, fruit and vegetables, and super - market items. Given that SVF is an important component of many cognitiv e assessment batteries, online collection of SVF data is highly useful in telehealth contexts. Native speakers of Turkish living in Türkiye or abroad were recruited through networks including Turk ish consulates. Demographic data was collected using Qualt rics, and SVF sequences were collected using a custom web app. 263 partic ipants completed the web survey, and 137 (52%) continued to the web app. We collected a total of n=311 SVF sequences (animals: n = 105, fruits/veg: n = 105, supermarket: n = 101). The online modality allowed us to reach 76 (72%) people from Türkiye and 29 (28%) from the Turkish diaspora. We discuss potential reasons for the drop off between survey and SVF data collection, as well as issues arising when collecting spoken SVF data.

Email: Maria Wolters, Maria. Wolters@ed.ac.uk

PSYCHONOMIC SOCIETY®
Follow us on Twitter! @Psychonomic_Soc
Tweet about your experience: #psynom23
A PSYCHONOMIC SOCIETY PUBLICATION
www.psychonomic.org

A.10 Turkish version of Troyer Taxonomy

Table A.3: Extended version of Troyer Taxonomy in Turkish

English main group name	Turkish main group name	English sub group name	Turkish sub group name	English sub group animals	Turkish sub group animals
Living Environment	Yaşam ortamı	Africa	Afrika	aardvark, antelope, buffalo, camel, chameleon, cheetah, chimpanzee, cobra, eland,elephant, gazelle, giraffe, gnu, gorilla, hippopotamus, hyena, impala, jackal, lemur, leopard,lion, manatee, mongoose, monkey, ostrich, panther, rhinoceros, tiger, wildebeest, warthog,zebra	yer domuzu, antilop, bizon, deve, bukalemun, çita, şempanze, kobra, afrika geyiği, fil, ceylan, zürafa, öküz başlı antilop, goril, su aygırı, sırtlan, impala, çakal, lemur, leopar,aslan, deniz ineği, firavun faresi, maymun, devekuşu, panter, gergedan, kaplan, afrika antilobu, yaban domuzu, zebra, pars, çıta, mirket, şebek, hipopotam, arslan, gepard
Living Environment	Yaşam ortamı	Australia	Avustralya	emu, kangaroo, kiwi, opossum, platypus, Tasmanian devil, wallaby, wombat	emu, kanguru, kivi, keseli sıçan, ornitorenk, tazmanya canavarı, valabi, vombat, dodo, koala, panda
Living Environment	Yaşam ortamı	Arctic/Far North	Kuzey Kutbu / Uzak Kuzey	auk, caribou, musk ox, penguin, polar bear, reindeer, seal	dalıcımartı, karibu, misk öküzü, penguen, kutup ayısı, ren geyiği, fok, mamut
Living Environment	Yaşam ortamı	Farm	Çiftlik	chicken, cow, donkey, ferret, goat, horse, mule, pig, sheep, turkey	tavuk, inek, eşek, gelincik, keçi, at, katır, domuz, koyun, hindi, boğa, buzağı, büyükbaş, bıldırcın, camız, civciv, dana, horoz, kaz, ördek, sıpa, sığır, manda, tosun, tay, oğlak, koç, kuzu, küçükbaş, hayvan
Living Environment	Yaşam ortamı	North America	Kuzey Amerika	badger, bear, beaver, bobcat, caribou, chipmunk, cougar, deer, elk, fox,moose, mountain lion, puma, rabbit, raccoon, skunk, squirrel, wolf	porsuk, ayı, kunduz, vaşak, karibu, gelengi, cougar, geyik, elk, tilki, moose, dağ aslanı, puma, tavşan, rakun, kokarca, sincap, kurt, boz ayı, dağkedisi, possum, kayoti
Living Environment	Yaşam ortamı	Water	Su	alligator, auk, beaver, crocodile, dolphin, fish, frog, lobster, manatee, muskrat, newt,octopus, otter, oyster, penguin, platypus, salamander, sea lion, seal, shark, toad, turtle, whale	aligator, dalıcımartı, kunduz, timsah, yunus, balık, kurbağa, ıstakoz, deniz ineği, misk faresi, semender, ahtapot, su samuru, istiridye, penguen, ornitorenk, Salamandra, Deniz aslanı, fok,köpek balığı, karakurbağası, kurbağa, balina, alg, büyükbeyazköpekbalığı, akya, carettacaretta, denizanası, denizatı, denizkaplumbağası, denizyıldızı, denizkestanesi, derekurbağası, kalamar, istakoz, karides, mürekkepbalığı, fokbalığı, sölenter, suyosunu, yosun, yengeç, sukaplumbağası
Human Use	insan kullanımı	Beasts of burden	Yük Hayvanları	camel, donkey, horse, llama, ox	deve, eşek, at, lama, öküz
Human Use	İnsan kullanımı	Fur	Kürk	beaver, chinchilla, fox, mink, rabbit	kunduz, çinçilla, tilki, vizon, tavşan
Human Use	İnsan kullanımı	Pets	Evcil Hayvanlar	budgie, canary, cat, dog, gerbil, golden retriever, guinea pig, hamster, parrot, rabbit	muhabbetkuşu, kanarya, kedi, köpek, gerbil, Golden Retriever, Gine domuzu, hamster, papağan, tavşan
Zoological Categories	Zoolojik Kategoriler	Bird	Kuş	budgie, condor, eagle, finch, kiwi, macaw, parrot, parakeet, pelican, penguin, robin,toucan, woodpecker	muhabbetkuşu, kondor, kartal, ispinoz, kivi, ara papağanı, papağan,muhabbetkuşu, pelikan, penguen, Kızılgerdan, tukan,ağaçkakan, akbaba, atmaca, baykuş, bülbül, bıldırcın, civciv, dodo, doğan, flamingo, güvercin, horoz, karabatak, keklik, kaz, karga, leylek, kumru, kuğu, kuş, sülün, turna, ördek, şahin, tavuskuşu, serçe, saksağan, martı, kırlangıç, yarasa, sümsükkuşu
Zoological Categories	Zoolojik Kategoriler	Bovine	Sığır	bison, buffalo, cow, musk ox, yak	bizon, yaban öküzü, inek, misk öküzü,Tibet sığırı, boğa, buzağı, büyükbaş, camız, dana, sığır, manda, tosun
Zoological Categories	Zoolojik Kategoriler	Canine	Köpek	coyote, dog, fox, hyena, jackal, wolf	kır kurdu, köpek, tilki, sırtlan, çakal, kurt, kayoti
Zoological Categories	Zoolojik Kategoriler	Deer	Geyik	antelope, caribou, eland, elk, gazelle, gnu, impala, moose, reindeer, wildebeest	antilop, karibu, afrika geyiği, elk, ceylan, öküz başlı antilop, impala, moose, ren geyiği, afrika antilobu, karaca, gazel
Zoological Categories	Zoolojik Kategoriler	Feline	Kedi	bobcat, cat, cheetah, cougar, jaguar, leopard, lion, lynx, mountain lion, ocelot,panther, puma, tiger	vaşak, kedi, çita, cougar, jaguar, leopar, aslan, vaşak,dağ aslanı, Oselo, panter,puma, kaplan, dağkedisi, vankedisi, pars, çıta, arslan, gepard
Zoological Categories	Zoolojik Kategoriler	Fish	Balık	bass, guppy, salmon, trout	levrek, Lepistes, somon, alabalık, akya, büyükbeyazköpekbalığı, istavrit, hamsi, kefal, lüfer, mezgit, palamut, sazan, pirana, sardalya, vantuzbalığı, uskumru, yılanbalığı, çinekop, çipura
Zoological Categories	Zoolojik Kategoriler	Insect	Böcek	ant, beetle, cockroach, flea, fly, praying mantis	karınca, böcek, hamamböceği, pire, Peygamber Devesi, akrep, arı, ağustosböceği, eşekarısı, güve, kelebek, kırkayak, sinek, sivrisinek, solucan, çekirge, tırtıl, uğurböceği, sümüklüböcek, tarantula, yabanarısı, örümcek, çiyan, ipekböceği, salyangoz, karafatma, ökenek
Zoological Categories	Zoolojik Kategoriler	Insectivores	Böcekçil	aardvark, anteater, hedgehog, mole, shrew	yer domuzu, Karınca yiyen, kirpi, köstebek, sivri fare
Zoological Categories	Zoolojik Kategoriler	Primate	Primat/ Maymun	ape, baboon, chimpanzee, gibbon, gorilla, human, lemur, marmoset, monkey,orangutan, shrew	ape, babun, şempanze, gibon, goril, insan, lemur, marmoset, maymun, orangutan, sivri fare, şebek, tembelhayvan
Zoological Categories	Zoolojik Kategoriler	Rabbit	Tavşan	Coney, hare, pika, rabbit	Coney, yabani tavşan, pika, tavşan
Zoological Categories	Zoolojik Kategoriler	Reptile/ Amphibian	Sürüngen/ Amfibi	alligator, chameleon, crocodile, frog, gecko, iguana, lizard, newt,salamander, snake, toad, tortoise, turtle	Aligator, bukalemun, timsah, kurbağa, geko, iguana, kertenkele, semender,Salamandra, yılan, karakurbağası, tosbağa, kaplumbağa, anakonda, carettacaretta, denizkaplumbağası, derekurbağası, dinozor, ejderha, kobrayılanı, köryılan, piton, sürüngen, sukaplumbağası, karakaplumbağası
Zoological Categories	Zoolojik Kategoriler	Rodent	Kemirgen	beaver, chinchilla, chipmunk, gerbil, gopher, groundhog, guinea pig, hamster,hedgehog, marmot, mole, mouse, muskrat, porcupine, rat, squirrel, woodchuck	kunduz, çinçilla, gelengi, gerbil, gopher, dağ sıçanı, gine domuzu, hamster, kirpi, marmot, köstebek, fare, misk faresi, oklu kirpi, sıçan, sincap, dağsıçanı, lağımfaresi
Zoological Categories	Zoolojik Kategoriler	Weasel	Gelincik	badger, ferret, marten, mink, mongoose, otter, polecat, skunk	porsuk, gelincik, sansar, vizon, firavunfaresi, su samuru, polecat, kokarca

A.11 Full table of Descriptive statistic of Word2Vec models of Turkish SVF analysis study

					Animal Category								Supermarket Category								Vegetables and Fruits Category							
Model name		hyperparameters			Number of switches				Mean Cluster size				Number of switches				Mean Cluster size				Number of switches				Mean Cluster size			
	Threshold	*d*	w	f	Mean	Mdn	max	min	Mean	Mdn	max	min	Mean	Mdn	max	min	Mean	Mdn	max	min	Mean	Mdn	max	min	Mean	Mdn	max	min
Word2vec	**0.75**	1000	10	cbow	16.74	16	45	5	1.43	1.38	2.5	1.1	17.5	17	31	4	1.42	1.35	2.36	1.06	16.6	17	39	3	1.47	1.45	2.2	1
		1000	**10**	**skipgram**	16.06	16	47	6	1.48	1.44	2.1	1	17.64	17	32	7	1.4	1.35	2.4	1	16.3	17	32	2	1.49	1.47	2.6	1.1
		1000	4	cbow	16.93	16	45	6	1.41	1.38	2.9	1.1	17.74	17	32	5	1.39	1.33	2.09	1.06	17.3	17	39	4	1.41	1.4	2	1
		1000	4	skipgram	16.48	16	48	6	1.44	1.4	2.3	1	17.74	18	35	6	1.39	1.35	2.36	1.07	17.7	18	31	3	1.38	1.35	2.1	1.1
		600	10	cbow	16.54	16	43	5	1.44	1.41	2.3	1.1	17.41	17	32	4	1.43	1.35	2.36	1.06	16.3	16	38	3	1.5	1.47	2.3	1
		600	**10**	**skipgram**	16.16	16	45	6	1.47	1.44	2.1	1.1	17.54	17	34	6	1.4	1.35	2.25	1.07	15.6	16	33	3	1.55	1.53	2.5	1.1
		600	4	cbow	17.45	17	45	6	1.37	1.33	2.2	1.1	17.6	17	32	4	1.41	1.35	2.17	1.06	16.9	17	38	3	1.45	1.45	2.2	1
		600	4	skipgram	16.35	16	47	6	1.46	1.42	2.3	1	17.6	17	34	6	1.4	1.35	2.18	1.07	17	17	32	3	1.43	1.4	2.2	1.1
		300	10	cbow	16.39	16	44	6	1.45	1.43	2.1	1.1	17.57	17	32	6	1.41	1.35	2.56	1	16.3	16	38	3	1.5	1.47	2.3	1
		300	10	skipgram	16.1	16	43	6	1.47	1.45	2.1	1.1	17.48	17	34	6	1.41	1.38	2.25	1.06	15.7	16	35	3	1.55	1.5	2.3	1.1
		300	4	cbow	17.06	16	45	6	1.4	1.37	2.2	1.1	17.57	17	32	4	1.41	1.37	2.36	1.06	17.4	17	39	4	1.4	1.39	2	1
		300	4	skipgram	16.29	16	46	5	1.47	1.42	2.7	1.1	17.79	17	34	6	1.38	1.35	2	1.06	16.9	17	33	3	1.43	1.41	2.1	1.1
	0.50	1000	10	cbow	11.19	10	34	3	2.16	2	4.1	1.4	11.56	11	23	2	2.2	2	6.5	1.2	10.1	10	25	0	2.56	2.25	19	1.3
		1000	10	skipgram	10.66	10	36	3	2.23	2.12	4.5	1.2	11.5	11	24	4	2.14	2	4.5	1.09	9.95	10	24	1	2.55	2.33	9.5	1.4
		1000	4	cbow	11.3	11	29	3	2.14	2.07	4	1.3	11.38	11	23	1	2.26	2.07	5.75	1.2	11	11	28	2	2.25	2.08	6.3	1.4
		1000	4	skipgram	11.06	11	33	3	2.16	2	4	1.3	11.74	12	23	4	2.12	2	4.6	1.09	10.9	11	24	1	2.27	2.17	5	1.2
		600	10	cbow	10.87	10	31	2	2.23	2.09	5	1.4	11.41	11	23	3	2.22	2.07	5.2	1.13	10.1	10	26	0	2.66	2.15	19	1.4
		600	10	skipgram	10.63	10	32	3	2.24	2.09	4.5	1.3	11.46	11	23	4	2.15	2	4.33	1.07	10.2	10	23	1	2.47	2.27	7	1.2
		600	4	cbow	11.2	11	28	3	2.15	2.07	4	1.3	11.54	11	24	1	2.22	2.1	5.75	1.13	10.6	10	27	2	2.31	2.1	4.8	1.4
		600	4	skipgram	10.82	10	34	4	2.21	2.09	4.5	1.2	11.61	11	24	4	2.16	2	5.2	1.09	11.3	11	25	2	2.16	2	4.7	1.3
		300	10	cbow	10.93	11	31	3	2.22	2.12	5	1.4	11.6	11	21	2	2.2	1.95	6.5	1.19	10.1	10	27	0	2.67	2.2	19	1.3
		300	10	skipgram	10.7	10	33	3	2.25	2.09	4.5	1.3	11.43	11	23	4	2.17	2.07	4.44	1.09	10.1	10	23	0	2.53	2.33	7	1.2
		300	4	cbow	11.31	11	27	3	2.14	2	4	1.3	11.68	11	23	1	2.22	2	5.75	1.2	10.7	10	25	1	2.35	2.14	7	1.4
		300	4	skipgram	10.87	10	34	3	2.21	2.12	4.1	1.2	11.56	11	22	4	2.15	2.06	5.2	1.09	10.4	10	23	1	2.36	2.25	4.2	1.2
	0.25	1000	10	cbow	5.68	6	18	0	4.59	3.75	20	1.8	5.6	6	12	0	4.92	3.88	21	1.5	4.26	4	13	0	6.04	5	26	2.3
		1000	10	skipgram	5.32	5	19	1	4.49	4	17	1.8	5.78	6	12	0	4.42	3.88	26	1.71	4.5	4	13	0	5.63	4.67	19	2
		1000	4	cbow	5.65	5	15	0	4.77	3.67	27	1.8	5.58	6	12	0	4.89	3.86	26	1.5	4.49	4	12	0	5.55	5	19	2
		1000	4	skipgram	5.63	5	20	0	4.5	3.71	27	1.8	5.83	6	12	1	4.36	3.78	15.5	1.33	4.61	4	14	0	5.33	4.6	19	1.8
		600	10	cbow	5.69	5	16	0	4.75	3.75	27	1.6	5.67	6	12	0	4.64	3.88	13	1.8	4.41	4	12	0	5.58	4.83	19	2.2
		600	10	skipgram	5.51	5	20	0	4.49	3.89	27	1.6	5.63	5	12	1	4.44	4	13.3	1.5	4.38	4	13	0	5.6	4.8	22	2.1
		600	4	cbow	5.63	5	16	0	5.07	3.62	27	1.6	5.73	6	12	0	4.69	4	26	1.46	4.49	4	15	0	5.52	5	19	1.8
		600	4	skipgram	5.7	5	19	0	4.28	3.67	14	1.8	5.96	6	13	1	4.23	3.71	13	1.5	4.61	4	12	0	5.4	4.67	19	1.6
		300	10	cbow	5.57	5	18	0	4.92	3.71	27	1.6	5.63	6	12	0	4.75	3.88	20	1.67	4.37	4	11	0	5.77	5	25	2
		300	10	skipgram	5.44	5	20	0	4.38	4	14	1.6	5.71	6	13	0	4.47	4	13.3	1.5	4.24	4	13	0	5.84	5	22	2
		300	4	cbow	5.61	5	15	0	4.83	3.75	21	1.6	5.71	6	11	0	4.72	3.7	26	1.5	4.5	4	12	0	5.63	4.67	25	1.8
		300	4	skipgram	5.82	6	18	0	4.29	3.6	17	1.6	5.88	6	12	1	4.26	3.8	13	1.5	4.77	5	12	0	5.16	4.4	19	1.8
Bigram					11.53	11	36	1	2.19	2.05	5.5	1.2	15.81	15	30	4	1.59	1.56	2.83	1.06	9.44	9	36	2	2.61	2.45	5.3	1.2
Troyer					12.08	12	38	3	1.97	1.91	3.8	1.1																

Table A.4: All descriptive statistics (Mean, Median(Mdn), Maximum(Max), Minimum(Min)) from 36 models of Word2Vec with Zeyrek Lemmatizer morphological analyzer. Scores are highlighted in green, fading from highest to lowest. Best performing models are bolded.