

Računalno potpomognuto usmjeravanje kod dvojezičnih govornika

Paun, Lukas

Doctoral thesis / Disertacija

2023

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Humanities and Social Sciences / Sveučilište u Zagrebu, Filozofski fakultet**

<https://doi.org/10.17234/diss.2023.130591>

Permanent link / Trajna poveznica: <https://um.nsk.hr/um:nbn:hr:131:679314>

Rights / Prava: [In copyright](#) / [Zaštićeno autorskim pravom.](#)

Download date / Datum preuzimanja: **2024-09-10**



Sveučilište u Zagrebu
Filozofski fakultet
University of Zagreb
Faculty of Humanities
and Social Sciences

Repository / Repozitorij:

[ODRAZ - open repository of the University of Zagreb
Faculty of Humanities and Social Sciences](#)





Sveučilište u Zagrebu

Filozofski fakultet

Lukas Paun

**RAČUNALNO POTPOMOGNUTO
USMJERAVANJE KOD DVOJEZIČNIH
GOVORNIKA**

DOKTORSKI RAD

Mentorica:
dr. sc. Maja Anđel, izv. prof.

Zagreb, 2023.



University of Zagreb

Faculty of Humanities and Social Sciences

Lukas Paun

COMPUTER-ASSISTED PRIMING IN BILINGUALS

DOCTORAL DISSERTATION

Supervisor:
Assoc. Prof. Maja Anđel, PhD

Zagreb, 2023

INFORMACIJE O MENTORICI

Maja Anđel profesorica je na Odsjeku za germanistiku Filozofskoga fakulteta Sveučilišta u Zagrebu. Ondje radi od 1997. godine prvo kao lektorica, a potom kao asistentica, docentica i naposljetku profesorica. Godine 2006. obranila je disertaciju pod naslovom "Strukturierung von Verbalargumenten im konnektionistischen Modell" pod mentorstvom prof. dr. sc. Zrinjke Glovacki-Bernardi s Odsjeka za germanistiku Filozofskoga fakulteta Sveučilišta u Zagrebu i doc. dr. sc. Georga Dorffnera s Medicinskoga sveučilišta u Beču. Od 2015. godine izvanredna je profesorica na Odsjeku za germanistiku Filozofskoga fakulteta Sveučilišta u Zagrebu. Do sada je bila članica Vijeća Filozofskoga fakulteta Sveučilišta u Zagrebu u razdoblju od 2010. do 2014., a između 2010. i 2012. bila je i pročelnica Odsjeka za germanistiku Filozofskoga fakulteta Sveučilišta u Zagrebu. Od 2014. predstojnica je Katedre za prevođenje njemačkoga jezika.

Bila je na studijskim boravcima u Austriji, Velikoj Britaniji, Kini i SAD-u. Također je bila mentorica poslijediplomskih studenata i poslijedoktoranada.

Područja su njezina znanstvena interesa psiholingvistika, kognitivna lingvistika, fonetika i fonologija njemačkoga jezika, sintaksa njemačkoga jezika i uvod u jezikoslovlje. Bila je članica predsjedništva Hrvatskoga DAAD Alumni kluba između 2005. i 2009. Od 2009. predsjednica je nekoliko Povjerenstava za ocjenu udžbenika njemačkoga jezika u osnovnom i srednjem školstvu pri Ministarstvu znanosti i obrazovanja Republike Hrvatske, a od 2014. članica je uredništva časopisa Zagreber Germanistische Beiträge.

SAŽETAK

Cilj je ovoga istraživanja otkriti predviđa li distribucijski semantički model (DSM) *word2vec* u izvedbi CBOW funkcioniranje mentalnoga leksikona dvojezične osobe. Osnovu analize činili su semantički prostori hrvatskoga i njemačkoga jezika (Günther, Dudschig i Kaup, 2015). Iz tih su se semantičkih prostora odabrale imenice, odnosno koncepti, a DSM je tražio njima najbližnje riječi. Valjanost se odluke DSM-a potom utvrdila u eksperimentu zadatkom leksičke odluke s usmjeravanjem (Meyer i Schvaneveldt, 1971). Ti su se rezultati usporedili s tradicionalno oblikovanim eksperimentom pomoću podataka iz Vezmar (2017) i De Deyne i sur. (2018).

Rezultati su se tih dvaju eksperimenata analizirali u okviru postojećih modela dvojezičnoga mentalnog leksikona kako bi se utvrdilo do koje mjere odabrani DSM prikazuje ljudsku jezičnu obradu. U oba je eksperimenta potvrđeno da se ciljne riječi koje su asocijativno povezane s usmjerivačima brže prepoznaju nego one koje to nisu. Ta je pojava u skladu s modelom BIA+ (van Heuven i Dijkstra, 2002) koji postulira da se brže prepoznaju one riječi koje su bliže u semantičkom prostoru. Pojava da su se ciljne riječi na J1 brže prepoznavale od onih na J2 može se protumačiti u okviru modela RHM (Kroll i Stewart, 1994). Prema razvojnoj komponenti toga modela iznimno vrsni govornici pristupaju konceptima izravno. Naposljetku se pokazalo da ne postoji razlika kod ispitnih parova riječi koji su generirani na tradicionalan način i pomoću modela *word2vec*. To je u skladu s radom Günthera, Rinaldija i Marellija (2019) po kojem DSM-ovi ne smiju nadmašiti rezultate koji postižu ljudi u rješavanju zadataka.

Razmatralo se također kako se rezultati dobiveni ovim istraživanjem mogu uključiti u izradu didaktičkoga materijala.

KLJUČNE RIJEČI

distribucijski semantički modeli, distribucionalizam, korpus, mentalni leksikon, poučavanje J2, psiholingvistika, usmjeravanje, utemeljenje, *word2vec*, zadatak leksičke odluke

ABSTRACT

This thesis investigates whether modern computer models can confirm how people encounter words and then use these findings in didactics. In recent years, computers have been used in psycholinguistics, but also in other areas of linguistics. Computer models in psycholinguistics were developed based on Harris's distributional hypothesis proposed in 1954. For this reason, such computer models are called distributional semantic models (DSM). According to Harris's hypothesis, words that have similar meanings share a similar context. Harris gave an example by comparing the words *oculist* and *eye-doctor*. These two words almost always occur in the same context and are interchangeable. They are synonyms. On the other hand, the word *lawyer* cannot be used in all contexts where the word *oculist* or *eye-doctor* is used, but only in some. Harris concludes this is due to frequency. In other words, whether it is possible to substitute the word *oculist* for the word *lawyer* in a certain context depends on how often both words occur in the same or a similar context.

While computer models used in computer science are being developed at a rapid pace, psycholinguistic research mainly uses models developed in the 1990s. Nevertheless, more modern computer models have recently been used in psycholinguistics. We have described the models LSA, HAL, BEAGLE and *word2vec* which are commonly used.

Landauer, Foltz and Laham (1998) defined the LSA model as an automatic mathematical and statistical technique used to make inferences and establish relationships between words in context. The context can be considered as a sentence, paragraph, or the entire text. LSA does not use dictionaries, knowledge, semantic networks, grammar and syntactic parsers or anything else created by humans. This model takes as input the raw text.

HAL is the second prominent distributional semantic model that Kevin Lund and Curt Burgess have worked on since 1992. HAL has many similarities with the earlier LSA model, but two major differences:

1. These two models have a different conception of context. In LSA, the elements of the co-occurrence matrix are units that form a cross-section of words and multi-sets (passage, part of speech, whole text, or anything that a researcher finds useful for his or her work). However, in HAL these elements are a cross-section of the co-occurrence of two words.
2. While in LSA all words within the same context are equally important, in HAL each word in the context has a different weighting factor, thus it is not equally important (Burgess, Livesay and Lund, 1998).

In contrast to these two models, Jones, Kintsch and Mewhort (2006) propose a different model that combines LSA and HAL. Their model stores the collected information in a composite holographic lexicon. The context includes discrete sentences, i.e. complete syntactic structures. They called their model BEAGLE, which is an abbreviation for Bound Encoding of the Aggregate Language.

Word2vec is a second generation model. It is not a unique monolithic algorithm. It consists of two separate models: the continuous skip-gram model and the continuous bag-of-words model (CBOW) (Mikolov et al., 2013a and 2013b). The earlier models (LSA, HAL and BEAGLE) are considered as count models, while *word2vec* is considered as a predictive model (Baroni, Dinu and Kruszewski, 2014). However, a common feature of *word2vec* and HAL is that both models use a window of a certain size around the central word that moves through the corpus (Mandera, Keuleers and Brysbaert, 2017).

CBOW aims to determine the central word using context words. On the other hand, the skip-gram model predicts the words before or after the central word at a certain distance from it.

Corpora are the basis for any DSM. They are the primary source of information on word distributions (Lenci, 2008). Any inferences that have been reached are entirely corpus-based. Therefore, in this study we have given a brief overview of what should be considered with a corpus in DSM research.

The mental lexicon is a “storehouse” of words with all important phonetic, phonological, morphological, syntactic and semantic features. Individuals actively use it every day so that the data it contains is always up-to-date. It grows constantly as the individual learns and adopts new words (Erdeljac, 2009).

There are two questions that researchers have addressed in studies of the bilingual mental lexicon. The first question was: Are the lexicons of the two languages that bilingual speakers speak separate or shared? In other words, are the units of the lexicon stored in two separate lexicons or in one shared lexicon? Many experiments have been conducted to understand and describe the organisation of the mental lexicon. In the meantime, researchers believe that the mental lexicon of a bilingual speaker is shared between two languages (de Groot, 2011).

The second question was: Is the access to the mental lexicon of bilingual speakers selective or non-selective, i.e. whether both lexicons are always active? Many experiments have led to the conclusion that the access to the mental lexicon is non-selective (de Groot, 2011).

We want to assess whether DSMs (more precisely, *word2vec*) resonate how the mental lexicon of a bilingual speaker functions. Hence, we have summarized different theories of the mental lexicon.

One major model of the bilingual mental lexicon is the Revised Hierarchical Model. It is based on the work of Kroll and Stewart (1994). They assume that in the early stages of L2 learning, a bilingual relies largely on L1 translations, i.e. on accessing the L2 meanings indirectly through lexical connections. As a bilingual becomes more proficient in L2, he or she increasingly relies on conceptual relations, even though the lexical ones have never been completely deactivated. Kroll and Stewart further emphasise that the lexical links between lexical units and languages and the conceptual links between separate languages and conceptual memory go both ways, but are different in strength.

Another important model of the bilingual mental lexicon is the Bilingual Interactive Activation Plus Model (BIA+) (van Heuven and Dijkstra, 2002). BIA+ extends the assumptions of the BIA model about orthographic representations to phonological and semantic representations. Orthographic representations excite phonological representations and these, in turn, excite representations at the next higher level and so on. However, that arousal depends on subjective frequency which suggests that recognition of phonological and semantic representations will occur later in L2 than in L1. Van Heuven and Dijkstra call this phenomenon the Temporal Delay Assumption. It has two consequences: (i) the interlingual effects will be greater in the direction from L1 to L2 than vice versa; and (ii) the absence of interlingual phonological and semantic effects may occur when the task allows a faster response to orthographic and other similar codes.

Many models of the mental lexicon have drawn conclusions from experiments with a primed lexical decision task. The key problem in previous research with this task was that the analysis of the results, based on prime and target lists generated by the researcher himself, was always done post hoc with DSM. Thus, only those studies that showed significant statistical effects were considered. We want to determine the predictive power of DSMs as one of the variables that can be manipulated during the experiment. Therefore, this study went in the opposite direction. Thus, we generated a list of prime words with *word2vec* first. The prime-target pairs were then tested in a primed lexical decision task focusing on German L2 speakers. We have chosen this type of task because it shows better effects than, for instance, the naming task (Lucas, 2000; Hutchison, 2003; Brysbaert et al., 2014).

This study was based on the following hypotheses:

H1: The reaction time of the targets offered as the most similar to the concepts according to *word2vec* should reflect the reaction time of the words collected through discrete free associations. This is consistent with the study by Günther, Rinaldi and Marelli (2019), who argue a DSM should not outperform humans in solving behavioural tasks.

H2: In the primed lexical decision task, participants will respond fastest for the words that are most similar to the prime. This is consistent with the BIA+ model. Similar words are less distant from each other and are therefore activated more quickly.

H3: The reaction time for German targets is shorter when they are preceded by a Croatian prime. This is consistent with the RHM model. Since bilingual speakers have a richer semantic representation of words in L1 than in L2, words in L1 are activated faster than in L2.

It is common for students to take part in language research, as they are the most accessible to researchers. Henry (2008) points out that this population is the most convenient for researchers in terms of money and time. Moreover, it is difficult to find enough participants who will spend time on such research. Nevertheless, we had to look for participants outside the usual places where participants are generated, like schools, universities, etc. because of the situation caused by the global COVID-19 pandemic.

We conducted the experiments using *OpenSesame* (Mathôt, Schreij, and Theeuwes, 2012), while *PsychoPy* (Peirce et al., 2019) was the backend. It measures the speed and accuracy of responses. We created two word lists for both experiments. Each list consisted of 80 pairs of prime and target words. The targets were pseudowords in half of the pairs. The pseudowords corresponded to the phonotactic and orthographic rules of Croatian and German, respectively. They were created with the programme *Wuggy* (Keuleers and Brysbaert, 2010). The other half consisted of pairs in which each target word was assigned to a prime which was semantically associated with the target word and another prime which was not semantically associated with the target word (Gulan, 2016; McNemara, 2005).

The first step in creating associative-semantically connected word pairs was to identify a concept. Each concept was a concrete noun. Subsequently, we identified the word that best corresponded to this concept in Croatian and German. In the first experiment, the associated primes were selected from a list of discrete associations collected by Vezmar (2017) and De Deyne et al. (2018), and in the second experiment, the associated primes were determined using *word2vec*.

54 participants took part in the first experiment. This experiment confirmed that the participants recognized semantically associated word pairs faster than pairs that were not associated. This is consistent with previous findings (Perea, Duñabeitia, and Carreiras, 2008;

Guasch et al., 2011; Bogunović and Čoso, 2018). BIA+ can explain this phenomenon. Since this model assumes that word meanings are associated with nodes in a shared semantic or lexical network, the recognition of a particular word activates a corresponding node in the network. The activation then automatically spreads further in the network by activating nearby nodes. The closer the nodes are, the faster they are recognized.

Regarding linguistic direction, target words seemed to be recognized more quickly when the prime was in German. Huynh and Witzel (2018) claim the current literature on semantically associated priming is inconclusive. Unbalanced bilinguals are assumed to have a richer semantic representation in L1 than in L2 (Duyck and Brybaert, 2004; Schoonbaert et al., 2009). This means the word in the participant's L1 activates more nearby conceptual nodes in the semantic network than a word in L2. Therefore, target words in L2 are recognized faster if an L1 prime precedes them. In contrast, balanced bilingual participants activate an equal number of close conceptual nodes in the semantic network (Guasch et al., 2011).

That target words in Croatian were recognized faster when they were preceded by a German prime can be explained by the ghosting effect (Finkbeiner, 2005) or the iconic persistence effect (Coltheart, 1980). This effect was caused by a blank screen of 250 ms occurring between the prime and the target word. This means the prime appears to be displayed longer than it actually is. Bilingual participants use this extra time to process the prime, which then affects the recognition of the target word (Huynh and Witzel, 2018). Hence, our participants had sufficient time to process the primes and therefore responded more quickly to L1 target words.

44 participants took part in the second experiment. We used the same target words as in the first experiment. The results confirmed the results of the first one. Participants recognized target words faster if an associated prime preceded. The results of the first experiment, in terms of the direction of priming, were also repeated. Participants recognized the target word faster when they were in their L1 and the primes in L2.

The cosine similarity between words proved to be relevant in determining the priming effect in the second experiment.

There was no statistical difference between the first and second experiment. This is consistent with the idea proposed by Günther, Rinaldi, and Marelli (2019). They argue DSMs must be able to complete tasks performed by humans and not outperform them. Sometimes DSMs achieve better results than humans (Mandera, Keuleers, and Brysbaert, 2017).

Finally, we have outlined how the insights gained from primed lexical decision tasks can be applied to foreign language methodology and didactics. We addressed how the findings

and results of previous studies relate to our findings. We have also summarized the results of this research into recommendations that can help teachers in their everyday work.

KEY WORDS

corpus, distributional semantic models, distributionalism, grounding, L2 teaching, lexical decision task, mental lexicon, psycholinguistics, priming, *word2vec*

Sadržaj

1. UVOD.....	1
2. ZNAČENJE.....	5
3. KRATAK PREGLED POVIJESTI DISTRIBUCIONALIZMA.....	9
4. DISTRIBUCIJSKI SEMANTIČKI MODELI (DSM).....	12
5. UTEMELJENJE.....	17
6. OPIS RADA POJEDINIH DSM-OVA.....	24
6.1. Latentna semantička analiza (LSA).....	24
6.2. Hiperprostor analogan jeziku (HAL).....	39
6.3. Vezano enkodiranje cjelokupne jezične okoline (BEAGLE).....	45
6.4. <i>Word2vec</i>	51
7. KORPUS.....	58
8. MODELI DVOJEZIČNOGA MENTALNOG LEKSIKONA.....	62
8.2. Inhibicijski kontrolni model (IC).....	64
8.2. Dvojezični interaktivni model leksičkoga pristupa i slušnoga prepoznavanja riječi (BIMOLA).....	67
8.3. Revidirani hijerarhijski model (RHM).....	69
8.4. Hipoteza o epizodičkom pamćenju.....	73
8.5. Bilingvalni interaktivni aktivacijski model plus (BIA+).....	74
8.6. Model distribucijskih konceptualnih obilježja (DCFM).....	80
8.7. Model značenja.....	81
9. CILJ ISTRAŽIVANJA.....	83
9.1. Hipoteze istraživanja.....	83
10. EKSPERIMENT 1: ZADATAK LEKSIČKE ODLUKE S USMJERAVANJEM S DISKRETNIM SLOBODNIM ASOCIJACIJAMA.....	84
10.1. Metoda.....	84
10.1.1. Ispitanici.....	84
10.1.2. Instrumenti i materijal.....	85
10.1.3. Postupak.....	88
10.2. Rezultati.....	90
10.3. Rasprava.....	92
11. EKSPERIMENT 2: ZADATAK LEKSIČKE ODLUKE S USMJERAVANJEM S ISPITNIM MATERIJALOM GENERIRANOGA POMOĆU DSM-A <i>word2vec</i>	97

11.1. Metoda.....	97
11.1.1. Ispitanici.....	97
11.1.2. Instrumenti i materijal.....	98
11.1.3. Postupak.....	101
11.2. Rezultati.....	102
11.3. Rasprava.....	104
12.USPOREDBA PRVOGA I DRUGOGA EKSPERIMENTA.....	110
12.1. Rezultati i rasprava.....	110
13.PEDAGOŠKE IMPLIKACIJE.....	114
13.1. Metoda poučavanja stranoga jezika.....	114
13.2. Uloga kognata i konkretnih riječi.....	117
13.3. Uloga distribucijskih semantičkih modela.....	119
13.4. Uloga dobi i okruženja.....	120
13.5. Utjecaj zadatka, aktivnosti i prestiža.....	121
13.6. Utjecaj udaljenosti između jezika.....	122
13.7. Utjecaj fonologije na učenje J2.....	123
13.8. Posljedice dvojezičnosti.....	124
13.9. Preporuke za poučavanje J2.....	125
14.ZAKLJUČAK.....	128
15.LITERATURA.....	130
16.PRILOZI.....	156
16.1. Upitnik.....	156
16.2. Sličice koje su ispitanici vidjeli tijekom eksperimenta.....	160
16.3. Popis parova ciljnih riječi i pripadajućih usmjerivača u oba eksperimenta.....	164
16.4. Popis tablica.....	171
16.5. Popis slika.....	172
17.ŽIVOTOPIS AUTORA.....	173

1. UVOD

U ovom ćemo radu pokušati ustanoviti mogu li suvremeni računalni modeli potvrditi način na koji ljudi pristupaju riječima kako bi ih se moglo primijeniti u didaktici. U posljednjim se godinama u psiholingvistici, ali i u drugim granama jezikoslovlja sve više rabe računala. Postoje oprečna mišljenja glede toga mogu li računalni modeli preslikati ljudsko rezoniranje ili služe samo kako bi se potvrdile određene hipoteze (Landauer, 2007; Glenberg i Robertson, 2000; Barsalou, 2008; Mandera, Keuleers i Brysbaert, 2017; Lake i Murphy, 2020).

Pokušat ćemo pomoću dva eksperimenta, kojima se vrlo često koristi u istraživanjima u psiholingvistici, kognitivnoj lingvistici i neuroznanostima, otkriti mogu li računalni modeli zrcaliti ljudsku obradu riječi. Riječi kojima se ljudi koriste smještene su u mentalnom leksikonu. Mentalni je leksikon spremnik riječi s fonetskim, fonološkim, morfološkim, sintaktičkim i semantičkim informacijama. Drugim riječima, istražuje se kako pojedinac može dohvatiti informacije iz mentalnog leksikona i kako se njima može koristiti.

Upravo će se računalni modeli koji se temelje na značenju riječi uzeti kao osnova za ovo istraživanje. Računalni su se modeli u psiholingvistici razvili na temelju Harrisove distribucijske hipoteze koju je postavio 1954. Zbog toga se takvi računalni modeli i zovu distribucijski semantički modeli (DSM). Prema Harrisovoj hipotezi riječi koje imaju slično značenje nalazit će se u sličnom kontekstu. Drugim riječima, pojedine će se riječi pojavljivati u različitim tekstovima, no riječi koje ju okružuju bit će uvijek slične.

Valja razlikovati računalne modele kojima se koristi u psiholingvističkim istraživanjima od onih kojima se koristi u istraživanjima iz računalstva. Dok se računalni modeli kojima se koristi u računalnoj znanosti razvijaju strelovitom brzinom, u psiholingvističkim se istraživanjima rabe uglavnom modeli koji su nastali u 1990-ima. U posljednje su se vrijeme i u psiholingvistici počeli upotrebljavati suvremeniji računalni modeli. Zbog toga smo se odlučili za model *word2vec* koji je nastao nedavno – 2013. Njegov potencijal još nije iskorišten u psiholingvistici.

Budući da su procesi obrade jezika koji se odvijaju u ljudskome umu vrlo kompleksni, razvili su se brojni modeli koji opisuju mentalni leksikon. Oni služe za pojednostavljen prikaz procesa obrade jezika i njegove produkcije. Ti se modeli razlikuju prema svojim temeljnim postavkama. Mogu se podijeliti u dvije velike skupine: modularni modeli i interaktivni

modeli. Modularni modeli polaze od toga da je ljudski um skup nezavisnih funkcionalnih modula i temelje se na serijskom pretraživanju mentalnoga leksikona. S druge strane, interaktivni modeli polaze od distribuiranoga pristupa pri čemu važnu ulogu igra jezično i izvanjezično znanje (Cergol, 2011; Bogunović i Ćoso, 2019; Erdeljac, 2009). Modeli koji opisuju dvojezični mentalni leksikon nastali su na temelju jednojezičnih modela.

Povećanjem broja jezika usložnjavaju se i pretpostavke o organizaciji i funkcioniranju mentalnoga leksikona te način kako mu pristupiti. Postavlja se pitanje jesu li leksikoni svih jezika kojima se služi pojedinac međusobno odvojeni ili postoji samo jedan leksikon u kojem se nalaze leksički unosi iz svih tih jezika (Kroll i Stewart, 1994; Perani i sur., 1998; Marian, Spivey i Hirsch, 2003). Nadalje se postavlja pitanje kako im pojedinac može pristupiti: selektivno ili neselektivno (Gerard i Scarborough, 1989; van Heuven, Dijkstra i Grainger, 1998; Brysbaert i Dijkstra, 2006; Dijkstra i van Heuven, 2012). Drugim riječima, prvi smatraju da se leksički unosi aktiviraju samo u jednom jeziku, dok su blokirani u ostalim jezicima, a potonji smatraju da su leksički unosi istodobno aktivirani u svim jezicima.

Prema našim saznanjima do sada nije provedeno mnogo istraživanja koja su se koristila DSM-ovima kako bi se utvrdio način funkcioniranja i organizacije mentalnoga leksikona kod dvo- i višejezičnih govornika. Većina istraživanja koja se bave mentalnim leksikom takvih osoba polaze od toga da je njihov mentalni leksikon organiziran kao jedna cjelina, tj. kao jedan zajednički dvo- ili višejezičan leksikon (Perani i sur., 1998; Marian, Spivey i Hirsch, 2003; Berken i sur., 2015). U ovom će se radu nastojati ustanoviti potvrđuje li odabrani DSM rezultate dobivene u eksperimentima provedenima s višejezičnim ispitanicima. Pokaže li se DSM dobrim prediktorom ljudskoga jezičnog ponašanja, to može imati posljedice na buduće konstruiranje eksperimentalnih zadataka, ali i na primjenu dobivenih rezultata. Jedna je od mogućih primjena DSM-a u pedagoške svrhe, što će se pokazati u ovom radu.

Cilj je ovoga rada utvrditi sličnosti i razlike između rezultata dobivenih na konvencionalan način, tj. kada istraživač odredi usmjerivač i ciljnu riječ i rezultata dobivenih tako da se usmjerivač odredi pomoću DSM-a *word2vec*. Svrha bi, prema tome, bila kako dobivene rezultate primijeniti u izradi didaktičkih pomagala i sredstava te dati upute za rad s učenicima koji žele naučiti strani jezik. Pri tome će se uzeti u obzir i dosadašnja saznanja stečena istraživanjima koja su se koristila zadatkom leksičke odluke s usmjeravanjem.

Za cjelovit je računalni model značenja potrebno izgraditi konceptualnu strukturu svijeta u kojem ljudi žive, a time je povezano i značenje. Stoga će se u sljedećem poglavlju opisati kako se gledalo na značenje. Utvrdit će se kako je teško odrediti značenje riječi i kojim se to metodama pokušalo učiniti.

Treće se poglavlje bavi kratkim pregledom povijesti distribucionalizma, a četvrto će poglavlje dati pregled DSM-ova i opisati u koje se svrhe rabe.

U petom poglavlju osvrnut ćemo se na problem utemeljenja. Mnogi su znanstvenici zbog te pojave osporavali da DSM-ovi mogu biti modeli ljudske kognicije.

Tema je šestoga poglavlja način funkcioniranja četiriju modela DSM-a koja su našla širu primjenu u psiholingvističkim istraživanjima. Dat će se njihova matematička osnova i koji su koraci nužni u generiranju semantičkoga prostora s kojima će se dalje raditi.

Korpusi su važan dio DSM-ova jer se na temelju njih oblikuju semantički prostori. Stoga je sedmo poglavlje posvećeno njima. Utvrdit će se na koje se parametre mora paziti prilikom odabira tekstova za valjan korpus.

Osmo će se poglavlje baviti modelima dvojezične obrade jezika. Osvrnut ćemo se na to koji modeli jezične obrade postoje i objasniti im osnovna načela. U raspravi ćemo pokušati s njima objasniti rezultate dobivene eksperimentima.

U devetom će se poglavlju iznijeti što su ciljevi ovoga istraživanja i kojim se hipotezama vodimo.

Deseto je poglavlje predviđeno za opis prvoga eksperimenta. Opisat će se način određivanja parova usmjerivača i ciljne riječi te ispitanici. Također će se opisati eksperimentalni zadatak koji ispitanici trebaju obaviti.

U 11. će se poglavlju predstaviti drugi eksperiment. Struktura je ista onoj u prethodnom poglavlju. Ispitanici će ponovno biti podvrgnuti zadatku leksičke odluke s usmjeravanjem. Popis ciljnih riječi bit će jednak prvom eksperimentu, samo što će se ovoga puta usmjerivači odrediti pomoću računalnoga modela *word2vec*. Za to je bitan odabir korpusa.

U 12. će se poglavlju usporediti oba eksperimenta.

Trinaesto je poglavlje predviđeno za opis pedagoških implikacija, tj. kako se rezultati dobiveni ovim istraživanjem mogu uklopiti u dosadašnja istraživanja u metodici i didaktici poučavanja stranoga jezika. Osvrnut ćemo se na druga istraživanja koja su se služila zadatkom leksičke odluke s usmjeravanjem s ispitanicima različite vrsnosti. Na temelju toga

donijet ćemo preporuke kojima bi se trebalo voditi u pedagoškom radu te preporuke za izradu didaktičkih materijala kako bi učenici stranoga jezika brže i jednostavnije proširili vokabular.

Rad završava s 14. poglavljem koje ujedno predstavlja i zaključak.

2. ZNAČENJE

Jezik se razvio kako bi ljudi međusobno prenijeli različite informacije koje im mogu poboljšati život ili ga u ekstremnim situacijama spasiti. Te informacije moraju biti povezane sa svijetom koji čovjeka okružuje. Lake i Murphy (2020) smatraju da konceptualni pristup značenju povezuje riječi preko koncepata sa svijetom. Međutim, pojedini su znanstvenici osporavali da DSM-ovi uopće mogu dohvatiti značenje koncepata jer se ono konstruira iz iskustvenoga doticaja s predmetima na koje se odnose (Barsalou, 2008; French i Jacquet, 2004). Zbog toga ćemo u nastavku pokušati rasvijetliti kako se na značenje gledalo i kako mu se pristupalo. Težište ćemo postaviti na distribucionalističko stajalište.

Značenje nije lako definirati zbog čega su ga neki poput Leonarda Bloomfielda isključili iz istraživanja jezika (Lenci, 2008; Rumshisky, Verhagen i Moszkowicz, 2009; Sahlgren, 2008). Tako u svojoj knjizi „Language” (1933) ističe da je značenje slaba točka jezikoslovlja te da ljudsko znanje treba još mnogo proširiti i razviti kako bi se stekli uvjeti da bi se uopće započelo s istraživanjem značenja. Prema tome, u određivanju značenja treba uzeti u obzir izvanjezičnu stvarnost. Njegov sljedbenik Harris uviđa da izvanjezični čimbenici utječu na jezik i značenje, no unatoč tome značenje će se uvijek moći, barem jednim svojim dijelom, odrediti iz suodnosa prema drugim elementima sustava (Harris, 1968). To znači da će se značenje riječi moći odrediti na temelju toga s kojim se drugim riječima pojedina riječ najčešće pojavljuje. Riječi koje se češće pojavljuju u istom okruženju imat će ista ili slična značenjska obilježja (Firth, 1957). To je bitno za računalne modele kojima ćemo se baviti u ovom radu. Međutim, naš cilj nije definirati značenje pojedine riječi, već kako se mogu uspostaviti odnosi između riječi koje imaju određeno značenje i kakvi su ti odnosi. Neki istraživači u novije vrijeme smatraju da računalni modeli koji se temelje samo na analizi tekstova uopće ne mogu zahvatiti značenje riječi (Bender i Koller, 2020; Lake i Murphy, 2020).

Prema kognitivnoj lingvistici značenje je smješteno u umu svakoga govornika. On je u stanju proizvesti i razumjeti iskaze. Nadalje značenje nastaje u socijalnoj interakciji i nije u cijelosti određeno. To znači da mora postojati nešto konvencionalizirano oko čega se svi slažu kao temelj za daljnje određivanje značenja riječi (Langacker, 2008).

Leksemi s kojima se često susrećemo imaju nerijetko više značenja. Langacker (2008) ističe da takve višeznačne riječi imaju više međusobno povezanih značenja. Oni su

konvencionalizirani do određene razine. Neka su značenja višeznačne riječi prototipičnija od drugih. U našem ćemo se radu usredotočiti samo na takva prototipska značenja riječi.

Svaki pojedinac u svakodnevnom životu nema većih poteškoća u upotrebi pojedinih riječi, no čim se pred njega postavi zadatak da definira značenje neke riječi, nailazi na veće ili manje poteškoće. Relativno je lagano definirati značenje punoznačnih riječi poput konkretne imenice *stolac*. No i tu se već mogu pojaviti nedoumice. Spomenuta riječ može imati više značenja. Pogleda li se u „Hrvatski enciklopedijski rječnik” osnovno je značenje riječi *stolac* „komad pokućstva s naslonom, namijenjen sjedenju za jednu osobu“ (Anić i sur., 2003: 1241). Ako netko u trgovini namještaja nađe primjerak nekoga dizajnera koji ne odgovara prethodnoj definiciji, svejedno će znati je li dotični predmet stolac ili nije. Ljudi vrlo vjerojatno ne znaju točnu definiciju nekoga predmeta ili pojma kako je navedena u rječniku, međutim znaju ono što je ključno za pojedini predmet ili pojam. Prema tome, u stanju su upotrijebiti riječ za taj predmet ili pojam na valjan način u svakodnevnoj komunikaciji. Pretpostavlja se, dakle, da ljudi uče značenje pojedinih riječi prema tome kako ih se upotrebljava u raznim kontekstima, a ne učeći definicije (Miller i Charles, 1991).

Evert i Lenci (2009) u svojem radu određuju tri pristupa kako se značenje može utvrditi. Prvi postulira da je značenje riječi *stolac* skup svih stolaca koji postoje u ovome svijetu ili funkcija mogućih riječi na skupu stolac. Taj je pristup formalan odnosno logičko-matematički. Drugi pristup smješta značenje u naše glave odnosno um. Prema tom shvaćanju značenje je riječi *stolac* pojam STOLAC kao mentalna reprezentacija kategorije stolac koju se može naći u našem mentalnom leksikonu. Taj je pristup obilježen kognitivnom psihologijom. Posljednji, treći pristup, jest distribucijski. Prema njemu značenje se riječi *stolac* može uopćiti iz svih jezičnih konteksta u kojima se riječ *stolac* upotrijebila.

Semantička se reprezentacija u distribucionalizmu razlikuje od kognitivne i formalne po tome što se temelji na kontekstu i ovisi o njemu. Drugim riječima, značenje se određuje pomoću jezičnoga konteksta. Nadalje semantička je reprezentacija distribuirana i dinamička. To znači da se značenje ustanovljava iz različitih odnosa koje riječi međusobno uspostavljaju. Promjenom se konteksta može promijeniti značenje. Naposljetku, semantička je reprezentacija kvantitativna i stupnjevita što znači da je značenje riječi utemeljeno na statističkoj distribuciji u različitim kontekstima. Prema tome, značenje se riječi neće samo razlikovati po tome u kojim se kontekstima pojavljuje već i koliko je značajan kontekst u kojem se pojavljuje (Lenci, 2008).

Navest ćemo jedan primjer kako iz konteksta možemo izvesti značenje pojedinih riječi. Prikazan je niz rečenica u Primjeru 1 koji je Lin (1998) preuzeo od Nide (1975: 167). Iz navedenih se rečenica može zaključiti da istaknuta riječ *tezgüino* označava alkoholno piće napravljeno od žitarica. Naime, riječi koje okružuju *tezgüino* često okružuju i druge riječi poput *pivo*, *votka* i sl.

- (1) A bottle of *tezgüino* is on the table.
Everyone likes *tezgüino*.
Tezgüino makes you drunk.
We make *tezgüino* out of corn.

Prema distribucijskom pristupu Evert i Lenci (2009) ističu kako značenje riječi ovisi o kontekstima u kojima se riječi upotrebljavaju. Isprva su računalni modeli samo djelomično i vrlo ograničeno bili u stanju uključiti izvanjezični kontekst (slike, predmeti, pokreti i sl.) u svoju obradu, no u posljednje vrijeme doživjeli su veliki napredak. U tim se modelima uglavnom upotrebljavaju vizualni materijali poput slika, ali sve češće i audio i video materijali (Bruni i sur., 2014; Chen i sur., 2020; Kiela, 2017). Takvi modeli uglavnom trebaju biti predtrenirani (*pretrained*) i fino ugođeni (*fine-tuned*) (Chen i sur., 2020; Lake i Murphy, 2020).

Konstrukcija značenja ili odnosa među značenjima pomoću DSM-a krije i poteškoće jer se u cijelosti temelji samo na podacima koji su dostupni računalnoj obradi, a to su prije svega tekstovi koji se nalaze u novinama, u knjigama, na internetskim stranicama i sl. Sahlgren (2006; 2008) ističe da je značenje riječi koje se može izlučiti iz konteksta samo odraz tekstova koji su obuhvaćeni korpusom, ali ne i značenje koje se krije u našem umu. Taj jaz pokušava premostiti Louwerse (2011) svojom hipotezom simboličke međuovisnosti (*symbol interdependency hypothesis*). Prema toj hipotezi značenje se može konstruirati iz jezičnih simbola, dakle riječi, i iz percepcijskih iskustava, dakle onih iskustava koje svaki pojedinac doživljava, a koje su izazvale riječi. Naše znanje o riječima nije posljedica samo percepcijskih iskustava već u velikom dijelu i statističke učestalosti pojavljivanja riječi. Drugim riječima, percepcijske su informacije već uključene u informacije koje dobivamo posredstvom jezika.

Louwerse (2011) je nizom istraživanja došao do zaključka da jezik obuhvaća utjelovljene odnose. Utvrdio je da oni uključuju, primjerice, zamjenu modaliteta, ponudu, ikoničnost i dr. Zamjena modaliteta označava pojavu da se dvije rečenice koje pripadaju istom

osjetilnom modalitetu (vizualnom, auditivnom i sl.) prepoznaju brže nego da jedna pripada drugom modalitetu. Odnos ponude označava pojavu prema kojoj nema značajne razlike u uporabi alternativnoga mogućeg predmeta za obavljanje određene radnje ako primarni predmet, kojim se određena radnja obavlja, nije dostupan (primjerice, ne postoji razlika u tome da se alternativno upotrijebi majica kako bi se osušio pojedinac nakon što iziđe iz vode). Louwerse ističe dalje kako su ikonički odnosi uključeni u jezik. Ikonički poredak u izrazu *od glave do pete* mnogo je češći nego obrnut poredak što potvrđuje činjenicu da ljudi uobičajeno svijet promatraju odozgo prema dolje. Čak i ako su ti odnosi prikazani vodoravno u nizu riječi, vrijeme je reakcije u zadatku jezične odluke u tom slučaju znatno brže nego da je obrnut poredak.

U literaturi koja obrađuje dvojezični mentalni leksikon razlikuje se mentalna reprezentacija pojma od leksema, odnosno značenjska reprezentacija riječi od pravopisne/pravogovorne reprezentacije riječi. Drugim riječima, značenje je preslikavanje riječi na konceptualnu strukturu. U višejezičnom okruženju to znači da je mentalna reprezentacija jednaka za sve jezike koje pojedinac govori. Prema tome, leksemi se razlikuju po jezicima tako da svaki jezik ima zaseban leksički unos u mentalnom leksikonu (Cergol, 2011; Dijkstra, 2003; Dijkstra i van Heuven, 2002; Kroll i Stewart, 1994; Van Heuven i sur., 2008).

3. KRATAK PREGLED POVIJESTI DISTRIBUCIONALIZMA

Uobičajeno se smatra da je distribucionalizam započeo Zelligom Harrisom 50-ih godina prošloga stoljeća distribucijskom hipotezom koja glasi:

If we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distribution of A and B are more different than the distributions of A and C. In other words, difference in meaning correlates with difference of distribution. (Harris, 1954: 156)

Da bi svoju ideju učinio razumljivijom, Harris je dao primjer uspoređujući *okulista* i *očnoga liječnika* (*oculist* i *eye-doctor*). Te se dvije riječi pojavljuju gotovo uvijek u istim kontekstima i međusobno su zamjenjive. One su, dakle, sinonimi. Nasuprot tome, riječ se *odvjetnik* neće moći upotrijebiti u svim kontekstima u kojima se koristi riječju *okulist* ili *očni liječnik*, već samo u nekim. Harris zaključuje da je uzrok tomu frekvencija. Drugim riječima, hoće li se u danom kontekstu moći zamijeniti riječ *okulist* riječju *odvjetnik*, ovisi o tome koliko se puta obje riječi pojavljuju u istom ili sličnom kontekstu.

Harrisova se ideja distribucionalizma razvila isprva na području fonetske analize, a kasnije se proširila na ostala područja lingvistike ne bi li postala općom metodologijom kojom bi se moglo analizirati pojavnosti na svim jezičnim razinama. U članku se „Distributional Structure” Harris (1954) prvo bavi fonetikom. Polazi od toga da su fonemi strukturirani prema određenim obilježjima koja tvore organiziran sustav. U članku se usredotočio na to kako opisati svaki jezik s obzirom na distribucijsku strukturu. To prema Harrisu znači da će se, polazeći od temeljne razine, opisati odnosi između pojedinih glasova pri čemu na taj proces neće utjecati čimbenici poput povijesti i značenja. U nastavku rada prelazi na značenje. Harris smatra da riječi nemaju jedno, centralno značenje, već da im se značenje mijenja. Značenje je po njemu društvena manifestacija i nedohvatna lingvističkoj teoriji (Sahlgren, 2008). Jezično je značenje razlikovno. To znači da je razlika u značenju posredovana razlikom u distribuciji (vidi gornji citat, Harris, 1954: 156). Umjesto o razlikama može se govoriti o sličnostima: ako je riječ A sličnija riječi C nego riječi B, onda će i distribucija riječi A i C biti sličnija nego distribucija riječi A i B. Sahlgren (2006; 2008) je uočio kako Harrisov rad ima čvrstu metodičku osnovu, no vrlo je slabo semantički utemeljen jer ne otkriva koje su relevantne distribucijske informacije niti za koje su značenjske razlike odgovorne distribucijske razlike.

Harris je bio pod de Saussureovim utjecajem. Naime, tvrdnju o različitosti značenja preuzeo je od njega. De Saussure je u svojem djelu „Tečaj opće lingvistike“ iz 1916. uveo pojam vrijednosti koji igra važnu ulogu u distribucijskoj hipotezi. Samo se po vrijednosti

jezični znakovi razlikuju jedan od drugoga. Da bi se jedna riječ razlikovala od druge, nije dovoljno da određenom nizu glasova ili slova odgovara određen pojam, potrebno ju je k tomu odijeliti od svih drugih riječi s kojima ju možemo usporediti. Stoga riječi dobivaju svoju vrijednost tako što stoje u opreci s drugim riječima (Kovačec, 2001). Te su razlike bitan element sustava i njegov sastavni dio. To znači da riječi ne mogu postojati izolirano od sustava. Pojedinačni, izolirani jezični znak ne može biti u opreci ni s čim. Potrebni su mu, dakle, drugi znakovi od kojih će se razlikovati (Sahlgren, 2008). De Saussure oprimjeruje to razlikom između francuske riječi *mouton* i engleske riječi *sheep*. Te se dvije riječi razlikuju po svojoj vrijednosti. Francuska riječ u engleskom pokriva dva značenja – *sheep* za životinju i *mutton* za meso te životinje. Ta razlika u francuskom ne postoji (De Saussure, 2000). Prema tome, de Saussureova ideja vrijednosti temelj je jezičnom značenju. Sahlgren (2008) u svojem radu dalje zaključuje da je pojam vrijednosti uz dihotomiju sintagmatskih i paradigmatskih odnosa osnova diferencijalne semantike. Nastavlja revidirajući Harrisovu distribucijsku hipotezu:

A distributional model accumulated from co-occurrence information contains syntagmatic relations between words, while a distributional model accumulated from information about shared neighbors contains paradigmatic relations between words. (Sahlgren, 2008: 40)

To znači da DSM-ovi koji su nastali prikupljanjem podataka koji se usporedno pojavljuju u određenom kontekstu tvore sintagmatske odnose, dok DSM-ovi koji su nastali prikupljanjem podataka od riječi koje se pojavljuju u istom kontekstu, ali ne zajedno, tvore paradigmatske odnose. Sintagmatske se odnose još naziva usporedno pojavljivanje prvoga reda, a paradigmatske usporedno pojavljivanje drugoga reda.

Otpribliže u isto doba Firth formulira svoje viđenje distribucionalizma rečenicom: „You shall know a word by the company it keeps“ (Firth, 1957: 179). Za razliku od Harrisa, koji je svojom distribucijskom hipotezom isticao semantičku sličnost, Firth naglašava važnost kolokacija u jeziku. Prema njegovu se mišljenju u istraživanju značenja treba usmjeriti na kombinacije riječi, tj. kontekst jer se samo pomoću konteksta može zaključiti što pojedina riječ znači. Na koje se značenje određene riječi misli, može se utvrditi kontekstom u kojem se pojavljuje.

Pojavom generativne gramatike, kako pojašnjava Lenci (2008), nestaje zanimanje za distribucijske modele koji se temelje na kontekstu. Povećavao se, međutim, interes za *i*-jezik i jezične kompetencije idealnoga govornika. Prvi je pojam uveo Chomsky u svojem radu „Knowlegde of Language“ (1986) kako bi pojam gramatika (*grammar*) učinio razumljivijim. Do toga se trenutka gramatikom smatrala generativna teorija i znanje koje govornik posjeduje

o određenom jeziku koji govori. *I*-jezik se u tom radu smatra stanjem uma pojedinca. *I* u *i*-jeziku znači individualan, interni i intenzionalan (Chomsky, 1986; Chomsky, 1995; Scholz, Pelletier i Pullum, 2020). Idealan je pak govornik osoba koja savršeno poznaje svoj jezik i nije podložna ograničenjima poput nedostatka pažnje, zanimanja ili slaboga pamćenja (Chomsky, 1965).

Jezičnom se kompetencijom idealnoga govornika pokušalo objasniti jezičnu strukturu, a to je obuhvaćalo i semantiku. Zbog toga su distribucije nastale na temelju korpusa odbačene kao pouzdan izvor jezičnih dokaza. Isto je tako matematički pojam vjerojatnosti također odbačen kao prikladan model gramatičkoga opisa jezika. Nasuprot tome, kognitivno jezikoslovlje ističe konceptualan pogled na semantiku. Prema njemu značenje je nekoga leksičkog izraza konceptualizacija entiteta ili situacije koju može izazvati (Lenci, 2008).

Američki psiholog Osgood došao je na ideju da se značenje riječi može prikazati kao točka u višedimenzionalnom prostoru. Proučavao je kako ispitanici pridijevaju vrijednosti značenju pojedinih riječi na ljestvici čiji su krajevi međusobno oprečni (npr. *sretan-tužan*, *tvrd-mekan* i sl.). Dakle, prema Osgoodu i sur. (1957) značenje se riječi može odrediti kao točka u euklidskom prostoru. Ta se sličnost u značenju između dviju riječi može ustanoviti mjerenjem udaljenosti među tim dvjema točkama. Salton (1971) je prvi razvio DSM za svoj sustav pretraživanja dokumenata (*informational retrieval*) pod nazivom SMART (Turney i Pantel, 2010). Uspjeh DSM-a u pretraživanju dokumenata potaknuo je druge istraživače da ih primijene u drugim područjima obrade prirodnoga jezika kao što je mjerenje sličnosti između značenja riječi, fraza i rečenica, strojno učenje i dr. U kognitivnoj se znanosti, posebice u teoriji prototipova, DSM-ovima koristi, primjerice, tako što se koncepti predstavljaju kao vektori, a kategorije, kojima koncepti mogu pripadati, kao skup vektora (Turney i Pantel, 2010). Landauer i sur. (2007) tvrde da su DSM-ovi poput LSA-a i HAL-a, o kojima će kasnije biti više riječi, empirijski i teoretski vjerodostojni u opisu nekih aspekata ljudske kognicije. Na tom je tragu i ovaj rad.

U današnje se vrijeme DSM-ovi rabe u različitim područjima. Primjenjuju se u jezikoslovlju, računalnoj znanosti i kognitivnoj znanosti. U jezikoslovlju se, primjerice, upotrebljavaju u korpusnoj i računalnoj lingvistici te u psiholingvistici (npr. proučavanje kako djeca usvajaju jezik, kako se konstruira mentalni leksikon i sl.). Osim toga primjenjuju se i u leksikografiji pri izradi rječnika.

4. DISTRIBUCIJSKI SEMANTIČKI MODELI (DSM)

Distribucijski semantički modeli još su poznati pod nazivima kontekstno-teoretski modeli, korpusno utemeljeni modeli, statistički modeli, vektorski semantički modeli, modeli leksičkoga prostora, modeli semantičkoga prostora, geometrijski modeli značenja ili model obrade vektora (Lenci, 2008). U tim se modelima kontekst rabi za određivanje značenja riječi u jednom njihovom dijelu. Postoji mnogo modela, a oni se razlikuju po tome kojim se računalnim tehnikama koriste, po semantičkim osobitostima prema kojima se distribuira tekst, određivanju konteksta i sl.

U svojem su radu Turney i Pantel (2010) ustanovili da se mnoge DSM-ove može ujediniti pod statističkom semantičkom hipotezom. Statističkom se obradom uporabe riječi kod ljudi želi otkriti što oni podrazumijevaju pod značenjem određene riječi. Značenje se riječi prema tome predstavlja kao vektor. Drugim riječima, vektore čine elementi značenjskih obilježja koji se sastoje od drugih riječi s kojima se usporedno pojavljuju. DSM-ovi prikazuju tako značenje višedimenzionalnih vektora smještenih u višedimenzionalnom semantičkom prostoru. To znači da se riječi koje mogu imati više značenja prikazuju u različitim kontekstima koji im određuju značenje. Vektore treba shvatiti kao točke ili usmjerene dužine u semantičkom prostoru koji su određeni svojim koordinatama odnosno smjerom. Pretpostavlja se ako su vektori u nekom korpusu tekstova slični, da su značenja riječi slična. Ta se općenita hipoteza može dalje raščlaniti na nekoliko specifičnijih hipoteza, a to su hipoteza o multiskupovima, distribucijska hipoteza, proširena distribucijska hipoteza i latentna relacijska hipoteza. Svima njima je zajedničko to što se značenje statistički određuje kontekstom u kojemu se pojavljuju.

Hipoteza o multiskupovima polazi od toga da frekvencija pojavljivanja neke riječi u dokumentu utječe na važnost dokumenta pri njihovu pretraživanju. Drugim riječima, koliko se često pojavljuje određena riječ u nekom dokumentu, toliko će važniji biti dotični dokument.

Distribucijska hipoteza počiva na Harrisovoj i Firthovoj ideji o tome da će se riječi koje imaju slično značenje pojavljivati u sličnom kontekstu.

Proširena distribucijska hipoteza, kako samo ime već govori, proširuje ideju koja se krije iza distribucijske hipoteze s pojedinačnih riječi na obrasce koji se usporedno pojavljuju u tekstu i imaju slično značenje. Lin i Pantel (2001) su u istraživanju aktivnih i pasivnih konstrukcija uočili da se subjekti i objekti u njima često međusobno usporedno pojavljuju. Iz

toga zaključuju da su aktivne i pasivne konstrukcije u kojima se pojavljuju ti subjekti i objekti međusobno slične.

Latentna relacijska hipoteza polazi od činjenice da parovi riječi koji se usporedno pojavljuju u sličnim konstrukcijama imaju slične semantičke odnose. Turney i Pantel (2010: 149) to opimjeruju odnosom između *staklara* i *stakla* u konstrukcijama kao što su "staklar upotrebljava staklo da..." ili "staklar je oblikovao staklo u...". *Staklar* se može zamijeniti bilo kojom drugom riječi koja označava nečije zanimanje, dok je *staklo* materijal kojim se ta osoba uobičajeno služi. Oni tvrde da se iz takvih konstrukcija može vidjeti semantička povezanost između dviju riječi.

U nastavku ćemo rada detaljnije promotriti distribucijsku hipotezu. S obzirom na njezinu široku primjenu Lenci (2008) smatra potrebnim razlikovati slabu i jaku distribucijsku hipotezu.

Slabu distribucijsku hipotezu definira kao kvantitativnu metodu semantičke analize. Ova hipoteza razmatra usporedno pojavljivanje riječi. Na temelju toga se izvode zaključci o dubljim semantičkim vezama kojima se objašnjavaju površinski međusobni odnosi. Prema tome, ova hipoteza pokušava otkriti bitne aspekte značenja koje dijele riječi koje se često usporedno pojavljuju. Ovdje se pretpostavlja da postoji korelacija između značenjskoga sadržaja i leksičke distribucije. Slaba se distribucijska hipoteza zasniva na Harrisovoj ideji (1954). Isto je tako slaba verzija usporediva s istraživanjima u teorijskoj lingvistici. Tako je, primjerice, Levin (1993) klasificirala engleske glagole koristeći se distribucionalnim metodama. Nadalje se slaba distribucijska hipoteza koristila za automatsko prepoznavanje višerječnih izraza i kolokacija (Fazly i Stevenson, 2008), a Pustejovsky i Jezek (2008) pokazali su kako se ona može povezati sa semantičkom teorijom da bi se dobili uvjerljivi rezultati u području generativnoga leksikona.

Jaka pak distribucijska hipoteza prema Lenciju (2008) počiva na ideji Millera i Charlesa (1994). To je kognitivna hipoteza o obliku i podrijetlu semantičkih reprezentacija. Polazi se od toga da distribucija riječi u kontekstu ima važnu ulogu pri oblikovanju semantičkih reprezentacija. Jaka hipoteza pretpostavlja da distribucija riječi u kontekstu uzrokuje oblikovanje semantičke reprezentacije riječi. Prema ovoj hipotezi odnos između riječi i njezina konteksta ne upućuje samo na semantičko ponašanje riječi već i na mogućnost da se semantički sadržaj objasni na kognitivnoj razini. Učestalo pojavljivanje riječi u različitim kontekstima dovodi do toga da se stvori apstraktna kontekstna reprezentacija

najznačajnijih konteksta u kojima se pojedina riječ pojavljuje (Lenci, 2008). Drugim riječima, utvrđuje se u kojim se kontekstima riječ može pojaviti i stvara se mentalna predodžba o sadržaju te riječi. Jaku distribucijsku hipotezu potvrđuju i brojna istraživanja koja se bave proučavanjem različitih psiholoških fenomena kao što su ocjena sličnosti, semantičko i asocijativno usmjeravanje, slabljenje semantičke memorije i sl. (Jones i sur., 2006; Vigliocco i sur., 2004). Na tom su tragu i dva distribucijska semantička modela LSA i HAL. Njima se često koristi u kognitivnim i psihološkim istraživanjima.

Međutim, kritizirana je činjenica da se značenje riječi ne može odrediti samo pomoću statističkoga prebrojavanja usporednoga pojavljivanja u kontekstu. Sahlgren (2008) prigovora da DSM-ovi mogu odrediti značenje riječi samo iz onih tekstova koji su im služili kao korpus. Taj je prigovor pokušao otkloniti Louwerse (2011) ističući da bitan udio u značenju riječi čini jezični kontekst (kotekst), a ne samo izvanjezični kontekst. No osim toga pojavili su se i drugi prigovori i kritike koje se uglavnom tiču triju pojava: kompozicionalnosti, inferencije i referencije.

Kompozicionalnost predstavlja pojavu prema kojoj se značenje kompleksnih izraza može dokučiti na temelju značenja njezinih sastavnica. Kada se DSM-u prigovara da nije u stanju obuhvatiti semantičku kompozicionalnost, to znači da ne može odrediti značenje kompleksnih izraza. DSM, dakle, ne može utvrditi značenje složenih struktura iz značenja njezinih sastavnih dijelova. Već su Landauer i Dumais (1997) predložili jednostavno rješenje: vektori se pojedinih sastavnih dijelova složenoga izraza samo trebaju zbrojiti. Međutim, pokazalo se da to nije svrsishodno jer je njihov model komutativan što znači da je u kontekstu koji se analizira sasvim svejedno je li, primjerice, muškarac udario dijete ili obrnuto dijete muškarca. Model Landauera i Dumais, naime, ne uzima u obzir red riječi u rečenici iz kojega se mogu odrediti takvi odnosi. Da se razriješi taj problem, Jones i Mewhort (2007) napravili su model BEAGLE koji vodi računa o sintaktičkim odnosima. Isto su tako Baroni i Zamparelli (2010) istražili, primjerice, atributivne odnose između pridjeva i imenica. Kombinirali su imenice i atribuirane pridjeve. Njihov se pristup ne zasniva na neovisnoj distribuciji samih pridjeva, već na tome kako pridjevi modificiraju distribuciju imenica njihovom međusobnom kombinacijom. Na njihov se rad oslanjaju Grefenstette i sur. (2013) koji su svojim DSM-om pokazali dobre rezultate uspoređujući rečenice s prijelaznim i neprijelaznim glagolom po semantičkoj sličnosti. Socher i sur. (2011) su pak istražili parafraze rečenica. Svojim su modelom postigli najbolje rezultate do tada.

Pojam leksičke inferencije opisao je Marconi (1997) smatrajući ga sposobnošću ljudi da shvate odnose među riječima. Polazi se, naime, od toga da pojedina osoba zna da su, primjerice, psi životinje i da se treba kretati ako se nekamo želi stići. Problem leksičke inferencije Lenci (2008) vidi u tome da DSM-ovi ne mogu dokučiti da je *auto* hiponim *vozilu* iako dolaze do rezultata da su *auto* i *vozilo* međusobno slični. Drugačije postavljeno, riječ se *vozilo* može pojaviti u svim kontekstima u kojima se pojavljuje riječ *auto*, no riječ *auto* ne može se pojaviti u svim kontekstima u kojima se pojavljuje riječ *vozilo*. Slično je i s homonimima. To su riječi koje imaju isti oblik, ali drugačije značenje. Primjerice, riječ *list* ima sedam različitih značenja (1. organ kojime biljke dišu; 2. komad papira; 3. isprava; 4. novine; 5. razvučeno tijesto; 6. pismo; 7. stražnja strana potkoljenice; Anić i sur., 2003: 677). Kako bi se riješio taj problem, u međuvremenu su se razvili višeslojni računalni modeli koji mogu razlučiti između višeznačnoga konteksta i jednoznačnoga konteksta. Primjerice, u istraživanju Lakea i Murphyja (2020), u kojem su se koristili modelom BERT (Devlin i sur., 2019), pokazalo se da u višeznačnom kontekstu višeznačna engleska riječ *bank* ima značenje financijske institucije (*She sees the bank.*). U jednoznačnom kontekstu značenje riječi *bank* ovisi o tome kako je postavljen kontekst. Tako u rečenici *She swims to the bank* riječ *bank* sličnija je riječima koje označavaju pojas zemlje uz različite tekućice (*shore, beach*). Slično je u rečenici *She deposits it in the bank* gdje je reprezentacija riječi *bank* sličnija onim reprezentacijama koje potječu iz svijeta financija (*treasury, ATM*).

DSM-ovi su do stanovite mjere u stanju razlučiti inferenciju, no ne i referenciju. Referencija je pojava povezivanja riječi i izvanjezičnih pojava. DSM-ovi, dakle, ne mogu imenovati predmete i pojave te odrediti predmete i pojave u izvanjezičnom svijetu. Ovaj je prigovor DSM-u povezan s time što nisu utemeljene simboličke reprezentacije, već samo predstavljaju statističke odnose između simbola. Ovo je usko povezano s pretpostavkom da konceptualne reprezentacije moraju biti utemeljene u osjetilnom i motoričkom sustavu (Barsalou, 2008). DSM-ove su oštro napali Glenberg i Robertson (2000) tvrdeći da oni ne mogu otkriti značenja koja proizlaze iz novih situacija. U trima su eksperimentima nastojali pobiti tvrdnju da DSM-ovi mogu na primjeren način odražavati značenje kod ljudi. U prva su dva došli do zaključka da ljudi mogu razlikovati smislene rečenice od besmislenih (npr. osušiti noge ručnikom i osušiti noge naočalima) za razliku od LSA-a. Na temelju trećega eksperimenta zaključili su pak da ljudi mogu shvatiti značenje riječi kojima je pridjenuto novo značenje za razliku od DSM-a. U tom su se eksperimentu rabili glagoli koji su izvedeni od

imenica (npr. *mobitel* → *mobitelirati*). Ispitanici su morali ocijeniti smislenost i gramatičku točnost te parafrazirati glagol. Međutim, LSA nije razlikovao smislene od besmislenih rečenica. U međuvremenu su se razvili mnogi različiti računalni modeli među kojima su neurosimbolički modeli koji su sposobni naučiti koncepte (Mao i sur., 2019). Unatoč tome suvremenim računalnim modelima nedostaje i dalje sposobnost da imenuju nove pojave i stvari ili da ih razumiju (Lake i Muphy, 2020).

5. UTEMELJENJE

Problem utemeljenja jedan je od glavnih prigovora koji kritičari DSM-a iznose kako bi osporavali mogućnost da postanu valjani modeli ljudske kognicije. Prvo, međutim, valja razlučiti tri ključna pojma koji se često brkaju. To su utemeljenje (*grounding*), utjelovljenje (*embodiment*) i situiranost (*situatedness*). Pezzulo i sur. (2012) te su pojmove definirali i oprimjerali. Utemeljenje se odnosi na činjenicu kako je kognicija utemeljena na fizikalnim pojavama u svijetu koji svakoga pojedinca okružuje. To su, primjerice, gravitacija i drugi zakoni fizike. Zoran primjer utemeljenja brojčane kognicije jest pojava kada se manji brojevi asociraju sa stupcima koji su niži, dok veći brojevi sa stupcima koji su viši. Za razliku od utemeljenja utjelovljenje je određeno ograničenjima vlastitoga tijela. Tako se pokazalo da osobe koje pripadaju zapadnom kulturnom krugu češće počinju brojati od palca lijeve ruke, dok osobe koje pripadaju bliskoistočnom kulturnom krugu počinju brojati od maloga prsta desne ruke (Lindemann, Alipour i Fischer, 2011). Situiranost se pak odnosi na kontekstnu ovisnost kognitivnih procesa. To su u prvom redu razna ograničenja i uvjeti koje pojedini zadatak postavlja pred ispitanika. Loetscher i sur. (2008) u svojem su istraživanju ustanovili da se u zadatku imenovanja brojki na ljestvici od 1 do 30 češće izgovaraju brojke koje su u donjoj polovici ljestvice ako im je glava u neutralnom položaju i ako im je okrenuta ulijevo nego ako je okrenuta udesno. Polovici se ispitanika reklo da prostorno zamisle ljestvicu kao onu na ravnalu gdje su niže brojke na lijevom kraju, a više brojke na desnom kraju, dok druga polovica nije dobila nikakve upute. Ustanovilo se da su ispitanici koji su dobili upute više naginjali k tomu da izriču brojke u donjoj polovici ljestvice.

Jedan od prvih prigovora DSM-u iznio je Searle (1980), a nešto kasnije i Harnad (1990). Searle u svojem članku opisuje misaoni eksperiment poznat pod nazivom "kineska soba". Radi se o tome da je u sobi zatvoren pojedinac s ogromnim brojem dokumenata na kineskom jeziku. Ta osoba ne zna kineski tako da ne bi mogla razlikovati kineske od japanskih znakova te reći postoji li ikakva razlika među njima. Potom dobiva drugi niz dokumenata na kineskom jeziku s pravilima o tome kako su znakovi međusobno povezani. Pravila su na materinskom jeziku i osoba ih stoga dobro razumije. Ona je sada u stanju prepoznati kineske znakove prema njihovu obliku. Nakon toga osoba dobiva treći niz dokumenata s uputstvima na materinskom jeziku koja joj omogućuju da povezuje elemente trećeg niza dokumenata s onima prvih dvaju. Ako se prvi niz dokumenata zamijeni pojmom skripta (*script*), drugi niz dokumenata pričom (*story*), treći niz dokumenata upitom (*question*),

niz simbola koji pojedinac uzvratiti nakon upita, tj. odgovorom (*answer to the question*) te niz pravila na engleskom jeziku programom (*program*), možemo ovaj misaoni eksperiment usporediti s DSM-om. Searle zaključuje da u tom slučaju taj pojedinac nije ništa drugo, nego jedna vrsta računala koji izvodi računalne operacije s formalnim elementima. Prema tome, kako osoba ne razumije ni riječ kineskoga, tako ni DSM-ovi ne razumiju priče odnosno tekstove koji im služe kao korpus.

Na Searlea se nadovezao Harnad (1990) koji u svojem radu prikazuje problem utemeljenosti simbola. Razlikuje dvije inačice toga problema: tešku i nemoguću. Prema prvoj pojedinac treba naučiti drugi jezik, a prema drugoj treba naučiti svoj prvi jezik isključivo pomoću jednojezičnoga rječnika. Dotični će se pojedinac stalno vrtjeti u krug jer će neprestano nailaziti na nove simbole koje treba shvatiti pomoću starih. Harnad smatra da je to osnovni problem s kojim se suočavaju simbolički modeli uma. Zaključuje da postoji samo jedan smjer prema kojem se može naučiti značenje riječi, a taj je od percepcije do simbola, tj. odozdo prema gore.

Shapiro (2008) se slaže da se značenje pojedinih simbola ne može dokučiti na temelju drugih simbola, no odbija Searleov misaoni eksperiment smatrajući ga tautologijom. Suprotstavlja mu svoj primjer "grčke sobe". Problem se svodi na to da osoba iz Grčke i osoba iz nekoga drugog kraja svijeta imaju različite predodžbe o glavnom gradu Grčke Ateni premda misle na isti grad. Dok Grk o Ateni misli da ima pet slova (*Αθηνα*), Amerikanac poput Searlea misli da ih ima šest (*Athens*). Tako Grk i Amerikanac više ne dijele istu predodžbu o Ateni jer obojica misle na onu riječ kako se piše na njihovu jeziku. Smjesti li se Amerikanac u "grčku sobu", a Grk mu na komadu papira postavi pitanje napisano alfabetom, Amerikancu bi to bilo besmisleno jednako kao da je smješten u kinesku sobu. Amerikanac bi se koristio pravilima za oblikovanje grčkih riječi ne poznavajući ih. Shapiro stoga predlaže sljedeće rješenje: osobu bi se iz "grčke sobe" trebalo pustiti iz nje kako bi istražila Grčku. S vremenom će dotični pojedinac naučiti koju riječ treba izgovoriti u kojoj situaciji. Grčki alfabet ima, prema tome, značenje za Grke, no ne i za ostale ljude koji nisu upoznati s tim pismom.

Lenci (2008) prigovor o neutemeljenosti i nemogućnosti referencijalnoga povezivanja vidi kao dio šire rasprave o dvjema suprotstavljenim hipotezama. Prema jednoj su koncepti sastavni dio osjetilnoga i motoričkoga sustava, a prema drugoj su koncepti simbolički i apstraktni entiteti (*Embodied Cognition Hypothesis vs. Abstract Cognition Hypothesis*). Često se polazi samo od toga da koncepti sadržavaju osjetilna i motorička iskustva, dakle iskustva

vanjskoga svijeta. To znači da se ta iskustva aktiviraju kada trebamo pristupiti značenju pojedine riječi. Nasuprot tome, Barsalou (2008) je oštar zagovornik tzv. utemeljene kognicije (*grounded cognition*). Razvio je svoj pristup koji je nazvao percepcijski simbolički sustav (*Perceptual Symbol System – PSS*). To je spoj tradicionalnoga pristupa koji se temelji na amodalnim (nepercepcijskim) simbolima i utemeljenju. PSS počiva na simulacijama. To su ponovljene aktivacije percepcijskih, motornih i introspektivnih stanja koja je pojedinac stekao. Ona su posljedica raznih iskustvenih doživljaja i ključne su za reprezentaciju značenja. Barsalou (2008; 1999) smatra da su simulacije i utjelovljenje ključni u kognitivnoj obradi. Tvrdi da se često smatra da su rezultati dobiveni raznim statističkim metodama neosnovani u pogledu utemeljenja. Međutim, tome ne mora biti tako. Statistička je obrada bitna u pristupu dinamičkoga sustava. Taj pristup polazi od toga da nema stalnih reprezentacija u umu, već da postoji veći broj sustava koji uključuju percepciju, djelovanje i kogniciju. Svaki od tih sustava može biti u stanju kontinuiranih promjena. Tijekom učenja ta se stanja povezuju međusobno i s izvanjezičnim okruženjem kako bi postigli svoje ciljeve. Sâm Landauer (1999) u odgovoru na Barsalouov članak smatra da su Barsalou i Searle prerigorozni u svojem ustrajanju na strogom utemeljenju. Landauer navodi primjer kako Boyleov zakon nije plin, već da simulira neke osobitosti plina. Stoga ni LSA nije čovjek, ali može simulirati neke osobitosti ljudske kognicije. Zaključuje da posljedica simulacije kognicije nije samo nešto poput kognicije, već jest kognicija.

Shapiro (2008) se slaže sa Searleom da mišljenje nije puka manipulacija simbola. No to ne znači da manipulacija interpretiranih simbola ne može predstavljati mišljenje. Ako računalo samo manipulira neinterpretirane simbole, ne može znati o čemu razmišlja. Takvo razmatranje daje prednost utjelovljenoj kogniciji jer poništava nedostatke koje su simbolisti imali u svojoj teoriji. Shapiro smatra da su sljedeća dva prigovora pridonijela jačanju utjelovljene kognicije: (i) razumijevanje ne nastaje manipulacijom samih simbola i (ii) simboli stječu svoje značenje. Polazi se, naime, od toga da su svi kognitivni modeli, pa i oni koji su se tradicionalno smatrali amodalnima (logičko razmišljanje, numerička i jezična obrada) utemeljeni na tjelesnim, afektivnim, percepcijskim i motoričkim procesima (Pezzulo i sur., 2012). Simbolisti se slažu s time da se ne mogu razumjeti simboli koji sami po sebi nemaju nikakvo značenje. Stoga je bitno da postoji semantika koja pridijeva značenje pojedinim simbolima. Prema Shapiru (2008) ključno je kako su simboli međusobno povezani te kako su povezani sa svijetom. Tek bi se tada razumjelo što oni zapravo znače. Louwerse i

Jeuniaux (2008) u svojemu članku ističu da je razumijevanje istodobno i utjelovljeno i simboličko. Tvrde da su mnogi eksperimenti tako postavljeni da daju prednost utjelovljenom tumačenju rezultata. Navode nadalje kako DSM-ovi poput LSA-a mogu predvidjeti vremenske i prostorne odnose što se smatralo nemogućim (Louwerse, 2006).

Barsalou (1999) uvodi percepcijske simbole koji su sami po sebi utjelovljeni. To su rekonstrukcije koje su nastale kao posljedice simulacije organizma. No Shapiro (2008) smatra da se takvi simboli ne razlikuju mnogo od računalnih simbola. Razlikuju se samo po tome što percepcijski simboli imaju ugrađenu vezu utjelovljenja. Shapiro, naime, smatra da percepcijski i računalni simboli nisu slični stvarima koje predstavljaju. Iako percepcijski simboli mogu biti slični percepcijskim reprezentacijama iz kojih su nastali, ne predstavljaju ih. To znači da ni jedan od tih dvaju simbola nije slučajno povezan s objektom na koji se odnose.

Wilson i Golonka (2013) idu korak dalje i tvrde da jezik nije ništa drugo nego jedna vrsta manifestacije raznih fizikalnih pojava. Jedina je razlika u njihovoj pojavnj strukturi i značenju. Svaka upotreba jezika (npr. govor, pismo) stvara određenu pojavnu strukturu (npr. govor stvara akustičku strukturu). Takva struktura može poslužiti kao jezična informacija. Budući da se jezične informacije smatraju jednakima percepcijskim informacijama, mogu se istovjetno analizirati. Društvene konvencije određenoga okruženja pridjenuli su svakoj strukturi određeno značenje. Kako se konvencije mogu mijenjati, može se mijenjati i značenje. Stoga je struktura konvencionalna i prema tome podložna promjenama. Nasuprot tome, percepcija nije toliko podložna promjena. Zbog toga je jezik nepostojaniji od percepcije. Wilson i Golonka (2013) u tome vide razlog zašto bismo se više trebali usredotočiti na pitanje kako naučiti upotrebljavati jezične informacije i odgovarati na njih, a ne kako učiti značenje riječi.

Često se tvrdi da su apstraktni pojmovi utemeljeni u amodalnom sustavu. Pulvermüller (2013) zaključuje da su svi simbolički sustavi nedvojbeno utemeljeni na percepciji, pa stoga i apstraktni. Ističe da su apstraktne riječi utemeljene pomoću emocija. Djeca se, naime, poučavaju značenju apstraktnih naziva emocija tako što im odrasli tumače kako se ljudi ponašaju kada pokazuju određene emocije. Dijete se može poučiti apstraktnoj riječi *sreća* jer pokazuje ponašanje tipično za situaciju u kojoj je sretno na što mu odrasla osoba skreće pozornost. Na taj će način dijete naučiti da se u takvim i sličnim situacijama koristi riječju *sreća*. Nove riječi djeca uče tako što se one često pojavljuju u istom okruženju s predmetom

na koji se pokazuje ili koji se drži u rukama. U kasnijim se fazama nove riječi mogu učiti pomoću već poznatih riječi. Riječi koje se pojavljuju u različitim kontekstima imaju slabe veze između svojih oblika i predmeta ili djelovanja na koje se odnose. To su poglavito riječi ili morfemi koji imaju gramatičko značenje (Moseley, Kiefer i Pulvermüller, 2016; Pulvermüller, 2013). Elektrofiziološko i neurometaboličko slikanje aktivnosti mozga pokazalo je aktivaciju pojedinih dijelova mozga ovisno o tome koje riječi ispitanik obrađuje. Ako se radi o riječima koje su povezane s bojama i oblicima, aktiviraju se dijelovi mozga koji su zaduženi za obradu boja i oblika (Pulvermüller, 2013; Moseley, Kiefer i Pulvermüller, 2016). Pokazalo se nadalje da neke apstraktne idiomatske rečenice aktiviraju određena prefrontalna i anteriortemporalna područja jače od konkretnih rečenica. Pulvermüller (2013) takvu pojavu multimodalne aktivacije korteksa smatra naznakom neutjelovljenoga razumijevanja značenja apstraktnih idioma. To potkrepljuje Louwerseovu tezu prema kojoj simboli ne moraju nužno biti uvijek utemeljeni (Louwerse, 2007). Praktičnije je, naime, zapamtiti amodalne simbole kao takve (Louwerse i Jeuniaux, 2008).

U posljednje se vrijeme mnogo radi na povezivanju različitih percepcijskih iskustava u jedinstven DSM (Kiela, 2017; Bruni i sur., 2014; Baroni i Lenci, 2010; Chen i sur., 2019) da bi se izbjegao problem neutemeljenosti DSM-a. Nadalje nekoliko istraživača ističe da jezični obrasci koji se uzimaju u obzir u istraživanjima DSM-a odražavaju znanje koje su ljudi stekli percepcijom, odnosno da jezično i perceptivno znanje u velikoj mjeri koreliraju (Louwerse, 2011; Louwerse i Connell, 2011; Tillman i sur., 2012). To bi moguće dovelo do krivoga zaključka da je problem oko utemeljenosti neosnovan, tj. da je znanje koje je dobiveno iz korpusa tekstova i znanje koje je stečeno percepcijom istovjetno, što ne odgovara činjenicama. Kintsch (2008) ističe da znanje koje predstavljaju DSM-ovi nikada ne može biti istovjetno ljudskom, no to ne predstavlja ograničenje zbog kojega se takvim isključivo neutemeljenim simboličkim modelima ne bi koristilo u istraživanju. Niz je radova (Andrews i sur., 2009; Baroni i sur., 2010; Riordan i Jones, 2011) pokazao da istraživanja koja se isključivo oslanjaju na korpus tekstova obuhvaćaju drugačije semantičke odnose. Takvi DSM-ovi obuhvaćaju enciklopedijske, funkcionalne i diskurzno važne odlike značenja riječi, ali ne i konkretne. Primjerice, takvi DSM-ovi neće *bananu* opisati ističući njezinu žutu boju, već će prije navesti da je tropsko voće i jestiva. Kada bi se za usporedbu pitalo ispitanike da opišu *bananu*, prevladali bi percepcijski opisi, dakle oni koji opisuju *bananu* kao žuto voće. Da se to izbjegne, napravljeni su multimodalni semantički modeli koji povezuju različita

percepcijska iskustva. Jedino se, naime, na taj način može opravdano govoriti o tome da DSM-ovi zrcale način na koji ljudi stječu i upotrebljavaju jezik. Ako se pak ta činjenica ne uzme u obzir, opravdano je govoriti o neutemeljenosti DSM-ova u percepciji i oni time gube svoju psihološku vjerodostojnost (Bruni i sur., 2014). Međutim, valja napomenuti da su Landauer i Dumais (1997) u svojem radu istaknuli moguće rješenje problema neutemeljenosti DSM-ova i postizanje psihološke prihvatljivosti tako što bi se osim tekstova u matrice uključile i slike.

Isprva se krenulo s onime što je najpristupačnije, a to su slike (Bruni i sur., 2014; Feng i Lapata, 2010; Kiela i sur., 2014; Leong i Mihalcea, 2011). Istraživanja s vizualnim multimodalnim distribucijskim semantičkim modelima pokazala su da kod konkretnih koncepata spoj informacija dobivenih iz korpusa tekstova i slika uvelike može poboljšati rezultate dobivene DSM-om, dok u suprotnom, ako se radi o apstraktnim konceptima, spoj vizualnih i tekstualnih informacija pogoršava rezultate. To je pak u skladu, kako Kiela i sur. (2014) ističu, s teorijom dvostrukoga kodiranja (Paivio, 1990). Prema toj su teoriji konkretni koncepti zapisani u jezičnom i u percepcijskom kodu, dok su apstraktni koncepti isključivo zapisani u jezičnom kodu. Ta je činjenica značajna utoliko što je većina punoznačnih riječi u svakodnevnom jeziku apstraktna. Zbog toga multimodalni DSM-ovi uključuju mogućnost razlikovanja konkretnosti i apstraktnosti (Kiela i sur., 2014; Bruni i sur., 2014).

Douwea Kiele u svojim radovima zaključuje da u auditivno i olfaktivno relevantnim podskupovima podataka dolazi do velikoga poboljšanja u izvedbi. Isto je tako točnost rezultata dobivenih auditivnim multimodalnim DSM-ovima jednaka onima s isključivo jezičnim reprezentacijama (Kiela i sur., 2014; Kiela i Clark, 2015; Kiela i sur., 2015; Kiela, 2017). To, međutim, ne iznenađuje s obzirom na iznimno malen broj dostupnih podataka koji se mogu upotrebljavati u takvim istraživanjima u odnosu na broj jezičnih korpusa. Općenito je vrlo malo riječi u korpusu relevantno za istraživanja u sferi auditivne i olfaktivne percepcije.

Pezzulo i sur. (2012) zagovaraju interakcionizam u utemeljenim računalnim modelima. To znači da je kognicija povezana s osjetilnim i motoričkim djelovanjima te da nije unaprijed određena raznim algoritmima, već da proistječe iz samoga djelovanja. Drugim riječima, algoritmi zaduženi za čitanje, prepoznavanje riječi ili predmeta, rješavanje raznih problema i sl. trebali bi pohraniti percepcijske, motoričke i afektivne informacije koje dobivaju tijekom izvođenja raznih radnji. Utemeljeni računalni modeli bi stoga trebali samostalno usvajati utemeljene simbole međudjelovanjem osjetilnih i motoričkih procesa. Pezzulo i sur. (2012)

polaze od toga da utemeljeni kognitivni procesi i motoričko djelovanje imaju iste mogućnosti i ograničenja. Zbog toga je dobro poznavanje utemeljenih kognitivnih procesa potrebno da bi se napravili njihovi realistični modeli. Takvi kognitivni robotički modeli zrcale sve mogućnosti utjelovljenja. Isto su tako pogodni za istraživanje okoline i učenja prema kriteriju pokušaja i pogreške. Također se pomoću kognitivnih robotičkih modela može istražiti interakcija tijela, djelovanja i percepcije. Wilson i Golonka (2013) otkrili su, primjerice, kako roboti mogu izvesti kompleksno ponašanje (npr. pokreti), a da nisu unaprijed upoznati s okolinom u kojoj su smješteni.

U posljednje su vrijeme računalni modeli koji uključuju računalni vid (*computer vision*) mnogo napredovali u pogledu utemeljenja tekstualnih reprezentacija. Međutim, takvi su modeli i dalje veoma udaljeni od načina kako ljudi razumiju značenje (Lake i Murphy, 2020). Bisk i sur. (2020) misle da većina računalnih modela nije mnogo unapredovala od obrade korpusa tekstova. Smatraju da će se rezultati u budućnosti poboljšati unapređenjem hardvera, povećanjem količine podataka za obradu i smanjenjem troškova obrade, no ono neće biti tako skokovito kao u posljednje vrijeme. Svoju tvrdnju temelje na tome što ni GPT-3 (Brown i sur., 2020) kao jedan od najnovijih modela nije značajno poboljšao rezultate na skupu podataka (*dataset*) LAMBADA (Paperno i sur., 2016). Taj je skup podataka napravljen s ciljem da se zahvati ljudska intuicija. Model GPT-3 je tako za samo osam posto nadmašio svoje prethodnike (Bisk i sur., 2020; Brown i sur., 2020). Lake i Murphy (2020) ne žele osporiti veliko dostignuće koje je postignuto modelom GPT-3, no smatraju kako je još dalek put pred nama da računalni modeli mogu shvatiti značenje na onaj način kako to čine ljudi.

6. OPIS RADA POJEDINIHM DSM-OVA

U ovom će se dijelu opisati što stoji iza pojedinih DSM-ova i kako funkcioniraju. Drugim riječima, objasniti će se kojim se matematičkim i statističkim tehnikama i metodama koriste pojedini DSM-ovi. Opisat ćemo one modele kojima se najčešće koristi u psiholingvističkim istraživanjima. Neki su modeli prema shvaćanjima računalne znanosti zastarjeli (npr. LSA, HAL i sl.), no u psiholingvistici su još uvijek vrlo popularni što se može vidjeti po broju objavljenih radova u kojima se koristi tim modelima i koji ih citiraju (Lake i Murphy, 2020).

6.1. Latentna semantička analiza (LSA)

LSA je proizišao iz modela LSI (*latent semantic indexing*; latentno semantičko idenksiranje) (Deerwester i sur., 1990; Furnas i sur., 1988; Lenci, 2008). Kako većina riječi ima više značenja, to je predstavljalo velik problem u pretraživanju informacija (*information retrieval*). Problem se očituje u tome što se prilikom pretraživanja pronalaze dokumenti koji ne predstavljaju sadržaj koji korisnik želi. Često se događa da riječi kojima se koriste ljudi koji traže određene informacije ne odgovaraju indeksiranim riječima. Sustav, naime, ne može prepoznati na koji dio značenja riječi pojedinac misli kada upisuje traženu riječ. Deerwester i sur. (1990) prepoznali su da taj problem ima dvije strane. S jedne je to sinonimija, a s druge polisemija. Identificirali su tako tri čimbenika koji utječu na taj problem:

1. Prilikom postavljanja upita pojedinci se ne koriste onim pojmovima kojima se opisuje ili indeksira dokument. Ovaj se problem pokušao riješiti proširivanjem značenja pojmova ili konstrukcijom tezaurusa.
2. Ne postoje metode koje bi automatski prepoznale polisemiju. Tom se problemu htjelo doskočiti izradom kontroliranoga vokabulara i ljudskih posrednika između onih koji postavljaju pitanja i sustava.
3. Sustavi indeksiranja i pretraživanja informacija svaku pojavnicu obrađuju zasebno. Dakle, ne postoji razlika u tome pretražuju li se pojmovi koji se gotovo uvijek pojavljuju zajedno ili pojmovi koji se rijetko pojavljuju zajedno. Upravo zanemarivanje redundancije može izobličiti rezultate.

Cilj je LSI-a bio taj da nađe način da se otkrije na koji se pojam misli prilikom postavljanja upita ili koji je pojam pridjenut nekom dokumentu. Drugim riječima, želi se

otkriti koja se latentna semantika krije iza takvoga pridijevanja. Da bi to postigli, Deerwester i sur. (1990) iskoristili su implicitne strukture višega reda (latentne strukture) povezivanjem pojmova i dokumenata kako bi otkrili kojim su vezama povezani. Polazili su od toga da su pojmovi i dokumenti koji su međusobno tijesno povezani smješteni blizu jedan drugome u semantičkom prostoru. Da bi uspjeli u tome, na raspolaganju im je stajalo nekoliko modela koji su obuhvaćali hijerarhijsko, particijsko ili preklopno grupiranje, ultrametrična i dodatna stabla te višedimenzionalni modeli udaljenosti koji se temelje na faktorskoj analizi (Furnas i sur., 1988). U odabiru pravoga modela Furnas i sur. (1988) vodili su se sljedećim trima kriterijima:

1. Model je trebao biti dovoljno dobar da može predstaviti semantičku strukturu između pojmova i dokumenata.
2. Model treba istodobno i eksplicitno prikazivati pojmove i dokumente. Dotadašnji su modeli uglavnom prikazivali ili samo pojmove ili samo dokumente.
3. Budući da su htjeli model koji će moći raditi s nekoliko tisuća dokumenata, trebao im je model koji se može prilagoditi takvim uvjetima. Drugim riječima, trebali su model koji se u cijelosti automatski prilagođava matrici sastavljenoj od pojmova i dokumenata.

Kako bi zadovoljili sve kriterije odlučili su se za faktorsku analizu kvadratne matrice (*two-mode factor analysis*) koja se temelji na rastavu na singularne vrijednosti (*singular value decomposition* – SVD). Deerwester i sur. (1990) navode kako prijašnji pokušaji primjene modela faktorske analize nisu mogli razviti njezin pun potencijal zbog toga što takvi modeli zahtijevaju izrazito intenzivnu računalnu obradu. Osim toga prijašnji su se pokušaji ograničavali na vrlo malen broj dimenzija ili su se rezultati dobiveni analizom pretvarali u jednostavna binarna grupiranja (Borko i Bernick, 1963). Nadalje su se neki pokušaji primjene faktorske analize oslanjali na izrazito dugotrajno i naporno skupljanje podataka kroz, primjerice, ispitivanje ljudi o semantičkoj sličnosti (Ossorio, 1966).

Deerwester i sur. (1990) su nadvladali opisane probleme koristeći se SVD-om koji su kasnije preuzeli Landauer i Dumais (1997) jer se pokazalo da LSI smanjuje smetnje (*noise reduction*) smanjenjem dimenzionalnosti i sažima podatke uklanjanjem redundancije.

Landauer, Foltz i Laham (1998: 267) definiraju LSA na sljedeći način:

LSA is a fully automatic mathematical and statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse. It is not a traditional natural language processing of artificial intelligence program; it uses no humanly constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, morphologies, or the like, and it takes as its input

only raw text parsed into words defined as unique character strings and separated into meaningful passages such as sentences or paragraphs.

Nazivi "pojmovi" i "dokumenti" u LSI-u zamijenjeni su nazivima "riječi" i "konteksti" (ili ponegdje "različnice" i "dokumenti" [Martin i Berry, 2007]) u LSA-u. Konteksti mogu biti različite veličine. Mogu biti rečenice, skupovi od nekoliko rečenica, odlomci, sastavci ili jedan njegov dio. Što će se pak uzeti kao kontekst u pojedinom istraživanju, ovisi o samim istraživačima. Bitno je to da je dio diskursa koji služi tomu da bi se iz njega mogla izlučiti latentna semantička struktura. Kako bi se pak do nje došlo, ključna je "moćna matematička analiza" koja se krije iza LSA-a, a to je rastav na singularne vrijednosti (SVD) (Landauer, Foltz i Laham, 1998).

Tablica 1: Popis od nekoliko naslova članaka koji su korišteni kao primjer prikaza funkcioniranja LSA-a. Naslovi su poslužili kao egzemplarni korpus kako bi se oprimjerio način rada modela LSA. Kraticom LSA označeni su oni naslovi koji su vezani uz teorijski okvir ili primjenu latentne semantičke analize, dok su kraticom BL obilježeni oni naslovi članaka kojima je tema prepoznavanje i procesiranje riječi kod dvo- ili višejezičnih govornika. Kurzivom su u naslovima označene one riječi koje se pojavljuju najmanje dva puta.

	Autori	Naslov
LSA1	George W. Furnas, Scott Deerwester, Susan T. Dumais, Thomas K. Landauer, Richard Harshman, Lynn A. Streeter i Karen E. Lochbaum.	<i>Information Retrieval using a Singular Value Decomposition Model of Latent Semantic Structure</i>
LSA2	Michael B. W. Wolfe, M. E. Schreiner, Bob Rehder, Darrell Laham, Peter W. Foltz, Walter Kintsch i Thomas K. Landauer	<i>Learning from text: Matching readers and texts by latent semantic analysis</i>
LSA3	Yao-Ting Sung, Chia-Ning Liao, Tao-Hsing Chang, Chia-Lin Chen i Kuo-En Chang	The <i>effect</i> of online summary assessment and feedback system on the summary writing on 6th graders: The LSA-based technique
LSA4	Fritz Günther, Carolin Dudschig i Barbara Kaup	<i>Latent Semantic Analysis cosines as a cognitive similarity measure: Evidence from priming studies</i>

LSA5	Bob Rehder, Michael L. Littman, Susan Dumais i Thomas K. Landauer	Automatic 3- <i>Language Cross-Language Information Retrieval with Latent Semantic Indexing</i>
BL1	Annette M. B. de Groot i Rineke Keijzer	What Is Hard to <i>Learn</i> Is Easy to Forget: The Roles of <i>Word Concreteness</i> , <i>Cognate Status</i> , and <i>Word Frequency</i> in Foreign <i>Language Vocabulary Learning</i> and Forgetting
BL2	Pascal Brenders, Janet van Hell i Ton Dijkstra	<i>Word recognition</i> in child second <i>language</i> learners: <i>Evidence from cognates</i> and false friends
BL3	Kristin Lemhöfer, Ton Dijkstra i Marije Michel	Three <i>languages</i> , one ECHO: <i>Cognate effects</i> in trilingual <i>word recognition</i>
BL4	Annette M. B. de Groot, Philip Delmaar i Stephen Lupker	The processing of interlexical homographs in translation <i>recognition</i> and <i>lexical decision</i> : Support for non-selective <i>access</i> to <i>bilingual</i> memory
BL5	Ton Dijkstra	<i>Bilingual visual word recognition</i> and <i>lexical access</i>

Valja isto tako napomenuti kako se radi o modelu koji ne rabi nikakav ljudski input. Zaključke o međusobnoj sličnosti dviju riječi računalo izvodi potpuno samostalno oslanjajući se isključivo na sâm korpus tekstova. To znači da kontekst nije unaprijed obrađen računalo ili na bilo koji drugi način (Landauer i Dumais, 1997; Landauer, Foltz i Laham, 1998).

LSA se izvodi u nekoliko koraka. Prvi od njih je izrada matrice. Kako se uglavnom radi o velikom broju dokumenata odnosno kontekstā s isto tako velikim brojem riječi, nastat će vrlo velika pravokutna matrica pri čemu stupci predstavljaju kontekste, tj. dokumente, a redci riječi. Elementi matrice su frekvencije pojavljivanja pojedine riječi u pojedinom kontekstu. Naslovi u Tablici 1 služe kako bi se pokazalo koje matematičke operacije izvodi model LSA tijekom pojedinih etapa i ne predstavlja nikakav aktualan korpus. Dakle, naslovi su u Tablici 1 poslužili kao konteksti kako bi se izradila Tablica 2. Dobivena matrica još nije prošla računalnu obradu i rezultat je običnoga prebrojavanja koliko se puta određena riječ pojavljuje u pojedinom kontekstu. Kako bi se, međutim, lakše moglo raditi s matricom,

provedena je lematizacija, tj. svođenje riječi na leme odnosno njihove natukničke oblike (npr. množina *languages* > jednina *language* u tekstu BL3 u tablici 2 ili svođenje izvedenih oblika glagola na infinitivni oblik *learning* > *learn* u BL1 gdje su obje inačice zastupljene). Isto su tako uzete u obzir samo one riječi koje se pojavljuju najmanje dva puta u različitim kontekstima u ovom malom korpusu od deset naslova znanstvenih članaka. Članci su grupirani u dva tematska područja. Prvo je vezano uz teoriju i primjenu latentne semantičke analize. Ti su članci označeni kraticom LSA. Drugu skupinu čine naslovi članaka s temom koja obrađuje prepoznavanje i obradu riječi kod dvo- i višejezičnih govornika. Ti su članci označeni kraticom BL.

Tablica 2: *Matrica tipa (m×n) u kojoj redci (m) predstavljaju riječi koje se mogu nalaziti u naslovima znanstvenih članaka, a stupci (n) naslove znanstvenih članaka. Element matrice (α_{ij} ; $i=1, \dots, m$; $j=1, \dots, n$) čini frekvencija pojavljivanja pojedine riječi u pojedinom naslovu znanstvenoga teksta. Drugim riječima, elementi matrice su frekvencije usporednoga pojavljivanja prvoga reda (first-order co-occurrence) i nisu računalno obrađeni (raw co-occurrence). Međutim, elementi matrice su lematizirani. Lematizacija jest proces u kojem se pojavnice iz korpusa svode na njihove natukničke oblike, tj. različite se pojavnice (članovi iste paradigme) svode na zajedničku lemu.*

	LSA1	LSA2	LSA3	LSA4	LSA5	BL1	BL2	BL3	BL4	BL5
information	1	0	0	0	1	0	0	0	0	0
retrieval	1	0	0	0	1	0	0	0	0	0
evidence	0	0	0	1	0	0	1	0	0	0
latent	1	1	0	1	1	0	0	0	0	0
semantic	1	1	0	1	1	0	0	0	0	0
learn	0	1	0	0	0	2	0	0	0	0
effect	0	0	1	0	0	0	0	1	0	0
analysis	0	1	0	1	0	0	0	0	0	0
language	0	0	0	0	2	1	1	1	0	0
cognate	0	0	0	0	0	1	1	1	0	0
word	0	0	0	0	0	2	1	1	0	1
recognition	0	0	0	0	0	0	1	1	1	1
lexical	0	0	0	0	0	0	0	0	1	1
bilingual	0	0	0	0	0	0	0	0	1	1
access	0	0	0	0	0	0	0	0	1	1

Česta je pojava da je broj redaka (riječi) mnogo veći od broja stupaca ($m \gg n$). Budući da se radi o velikom broju kontekstā, većina će se riječi pojavljivati vrlo rijetko pa će tako nastati rijetka matrica (*sparse matrix*). To znači da će elementi matrice uglavnom biti nule.

Međutim, prije nego što se uopće može pristupiti "moćnoj" matematičkoj analizi SVD-u, računalno neobrađena matrica treba proći korak predobrade. Taj korak uključuje postavljanje težinskoga faktora (*weighting*) na svaki element matrice koji je različit od nule ($a_{ij} \neq 0$). Funkcija je težinskoga faktora ta da postavlja male težine na visokofrekventne riječi, dakle riječi koje se pojavljuju u mnogo kontekstā, a velike težine na riječi koje se pojavljuju u nekim, ali ne svim kontekstima. Proces predobrade u modelu LSA uključuje izračunavanje lokalnoga i globalnoga težinskoga faktora za svaki element a_{ij} različit od nule. Određivanjem lokalnoga težinskog faktora povećava se odnosno smanjuje važnost pojedine riječi u kontekstu, a određivanjem globalnoga težinskog faktora povećava se odnosno smanjuje važnost pojedine riječi u svim kontekstima. Lokalni težinski faktor povezan je, dakle, s frekvencijom pojavljivanja riječi u pojedinom kontekstu, dok je globalni težinski faktor povezan s frekvencijom pojavljivanja riječi u svim kontekstima korpusa.

Postoji nekoliko različitih funkcija kojima se može mjeriti lokalni (l_{ij}) i globalni težinski faktor (g_i). Neke uobičajene funkcije lokalnoga težinskog faktora (l_{ij}) definirane su na sljedeći način:

1. frekvencija riječi ($l_{ij} = \text{tf}_{ij}$, pri čemu je tf_{ij} broj pojavljivanja određene riječi u pojedinom dokumentu)
2. binarna frekvencije ($l_{ij} = 1$ ako riječ postoji, u suprotnom 0)
3. logaritamska funkcija ($l_{ij} = \log(\text{tf}_{ij} + 1)$)

Neke pak uobičajene funkcije globalnoga težinskog faktora (g_i) definirane su ovako:

1. normalan g ($g_i = \frac{1}{\sqrt{\sum_j \text{tf}_{ij}^2}}$)
2. globalna frekvencija - inverzna frekvencija konteksta g/Idf ($g_i = \frac{\text{gf}_i}{\text{df}_i}$, pri čemu je gf_i učestalost pojavljivanja riječi i u svim kontekstima korpusa, a df_i broj dokumenta u kojima se pojavljuje riječ i)
3. inverzna frekvencija konteksta Idf ($g_i = \log_2 \frac{n}{1 + \text{df}_i}$ pri čemu je n ukupan broj kontekstā u korpusu)

$$4. \text{ entropija } (g_i = 1 + \frac{\sum_j p_{ij} \log p_{ij}}{\log n} \text{ pri čemu je } p_{ij} = \frac{tf_{ij}}{gf_i}).$$

Kako bi se došlo do težinskoga faktora, moraju se pomnožiti lokalni i globalni težinski faktor za svaki element a_{ij} u matrici \mathbf{A} ($a_{ij} = l_{ij} \cdot g_i$). Dumais (1991) je u svojem istraživanju težinskoga faktora u LSI-u došla do zaključka da najbolje rezultate pokazuje umnožak logaritamske funkcije i entropije. Landauer, Foltz i Laham (1998) zaključuju da je takvo postavljanje težinskoga faktora bolje za predstavljanje određenoga konteksta jer ističe riječi koje nose stanovito značenje.

Nakon toga se može pristupiti rastavu matrice na njezine ortogonalne sastavnice. Prema Martin i Berryju (2007) najpopularniji je rastav na singularne vrijednosti (SVD). Pomoću SVD-a se istodobno mogu prikazati vektori za riječi i kontekste. Nadalje se pomoću njega može obuhvatiti semantička struktura u nekom korpusu kontekstā. Prilagodбом broja dimenzija može se utjecati na prikaz odnosa riječi i kontekstā u vektorskom prostoru.

Matrica \mathbf{A} tipa ($m \times n$) s rangom $\mathbf{A} = r$ definirana je na sljedeći način:

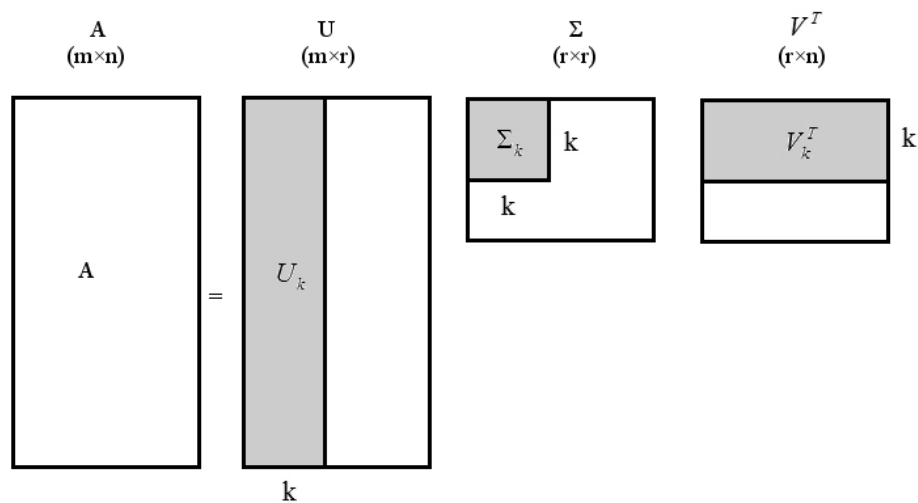
$$A = U \Sigma V^T$$

Pri tome su \mathbf{U} i \mathbf{V} ortogonalne matrice, a Σ dijagonalna matrica. Stupci matrice \mathbf{U} i \mathbf{V} su matrice lijevih i desnih singularnih vektora (*singular vector*), a matrica Σ matrica je singularnih vrijednosti (*singular value*) (Letsche i Berry, 1997). Drugim riječima, to bi u našoj Tablici 2 značilo da matrica \mathbf{U} predstavlja lijeve singularne vektore od deset dimenzija za 15 riječi, a matrica \mathbf{V} desne singularne vektore od deset dimenzija za deset kontekstā, dok je matrica Σ dijagonalna matrica od deset singularnih vrijednosti.

U okviru modela LSA značenje se može promatrati kao presjek značenja pojedine riječi u svim kontekstima u kojima se u danom korpusu pojavljuje. Češćim se pojavljivanjem riječi povećava dimenzionalnost matrice. Landauer, Foltz i Laham (1998) ističu da se smanjenjem dimenzionalnosti pokušava obuhvatiti međudjelovanje riječi i kontekstā te na taj način bolje opisati kognitivne veze u ljudskom umu. Drugim riječima, dimenzije koje su nastale smanjenjem dimenzionalnosti bile bi slične osnovnim semantičkim obilježjima pojedine riječi.

Za LSA je, dakle, ključno smanjenje dimenzionalnosti matrice \mathbf{A} . Smanjenjem se dimenzionalnosti otkrivaju semantički odnosi između kontekstā te se eliminiraju smetnje

(noise), tj. isključuju se one riječi koje ne pridonose razlikovanju kontekstā, utvrđuje se razlika u upotrebi riječi i sl. Smanjenjem broja stupaca matrice U i V te smanjenjem broja elemenata matrice Σ sa r na k nastaje nova matrica A_k tako da je $k \leq r$. Takvim se smanjenjem singularnih vektora i singularnih vrijednosti svi preostali elementi matricā do r izjednačuju s nulom. Ako se te tri derivirane matrice pomnože, nastat će matrica koja je približna originalnoj matrici A (Letsche i Berry, 1997; Martin i Berry, 2007). Shematski se to može prikazati kao na Slici 1.



Slika 1: Rastav na singularne vrijednosti matrice A tipa $(m \times n)$ ranga r na tri matrice ranga k $U_k \Sigma_k V_k^T$. Matrica A_k stvorena je tako da su se svi elementi osim prvih k elemenata ignorirali ili poništili, tj. izjednačili s nulom. Tako derivirana matrica A_k najbliža je originalnoj matrici A . Prvih k elemenata stupca matrice U i V međusobno su ortogonalni, dok prvih k elemenata reda matrice U i V to nisu. Matrica Σ je dijagonalna. Preuzeto od Martin i Berryja (2007).

Svaka je pravokutna matrica A_k sastavljena od triju drugih matrica U_k , V_k i Σ_k . Prva od njih, matrica U_k , ima jednak broj redaka kao originalna matrica, ali k stupaca. Ti su stupci derivirane varijable koji su linearno neovisni jedan o drugom. To znači da se iz jednoga stupca ne može linearnim kombinacijama stvoriti drugi stupac. Slično je s matricom V_k . Broj stupaca matrice V_k odgovara broju stupaca originalne matrice. Redci su matrice V_k također linearno neovisni jedan o drugome. Ti su k redci nastali deriviranjem singularnih vektora (Landauer i Dumais, 1997). Središnja je matrica Σ_k dijagonalna, tj. kvadratna matrica tipa $(k \times k)$. Elementi te matrice različiti su od nule samo uzduž središnje dijagonale matrice. Takve se derivirane konstante naziva singularnim vrijednostima. Njihova je zadaća povezati elemente preostalih

dviju matrica (Landauer i Dumais, 1997; Landauer, Foltz i Laham, 1998; Martin i Berry, 2007).

Kako je već ranije navedeno, upravo taj postupak smanjenja dimenzionalnosti matrice smanjuje redundanciju. Isto tako taj je postupak odgovoran za otkrivanje skrivenih semantičkih veza među kontekstima. Osim toga uzrok je tomu da se u semantičkom prostoru nađu u blizini slične riječi koje se nisu pojavljivale u istim kontekstima ili se konceptualno slični konteksti nalaze međusobno vrlo blizu premda ne dijele ni jednu riječ. Takvim se postupkom smanjenja dimenzionalnosti ne želi maknutim dimenzijama pridjenuti nikakvo značenje kako su to učinili Borko i Bernick (1963) kad su pokušali interpretirati što koja dimenzija znači. Pretpostavlja se da svaka dimenzija predstavlja jednu ili više semantičkih veza koje čine riječi i konteksti (Letsche i Berry, 1997).

Koju će pak vrijednost imati k , ovisi o samoj namjeni istraživanja i ciljevima koje se želi postići. U istraživanjima Landauera i Dumais (1997) pokazalo se da se najbolji rezultati postižu s dimenzionalnošću od oko 300. To su kasnije potvrdila i druga istraživanja (Berry i sur., 1999; Jessup i Martin, 2001; Lizza i Sartoretto, 2001; Steyvers, Shiffrin i Nelson, 2004). Cilj je naći optimalnu dimenzionalnost jer se na taj način optimalno komprimiraju podaci. Ako je dimenzionalnost premala, onda dobiveni podaci nisu dovoljno jasni, odnosno latentna se struktura ne odražava dovoljno detaljno. U suprotnom slučaju, ako je dimenzionalnost prevelika, ne mogu se ustanoviti međusobne semantičke veze (Steyvers, Shiffrin i Nelson, 2004).

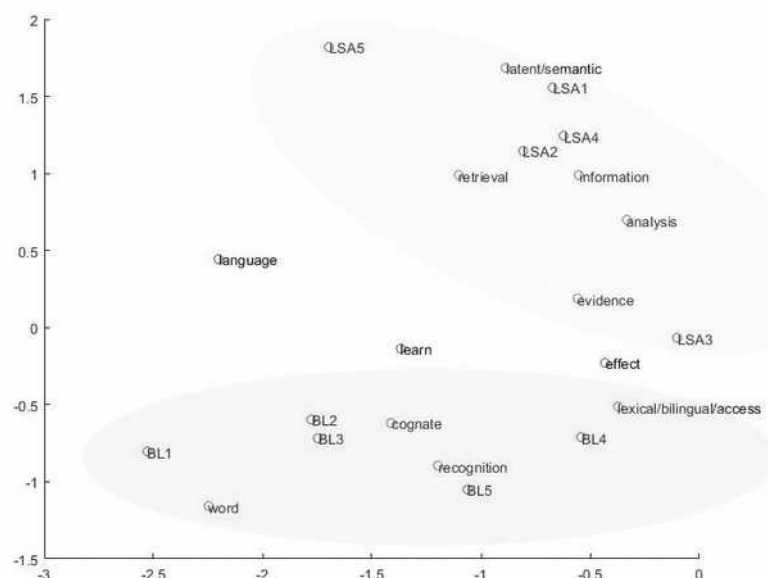
Budući da je LSA jedan od najpopularnijih DSM-ova, a SVD najčešća matematička analiza, odlučili smo se detaljnije objasniti faktorsku analizu koja se krije iza SVD-a. Pri tome smo se vodili prvenstveno radovima Landauera i Dumais (1997), Martin i Berryja (2007), Deerwestera i sur. (1990), Landauera, Foltza i Lahama (1998), Letschea i Berryja (1997) te Furnasa i sur. (1988) i primjerima iz njih, a više se o rastavu na singularne vrijednosti i o linearnoj algebri koja joj je temelj može saznati u radovima Stranga (2009), Goluba i van Loana (1996) te Horvatića (2004).

Tablica 3: Rastav na singularne vrijednosti matrice tipa (15×10) iz Tablice 2 gdje su elementi matrice frekvencije pojavljivanja određene riječi u pojedinom kontekstu. Zasjenčani su stupci vrijednosti prvih k dimenzija. SVD je napravljen pomoću MATLAB 2015b.

U=									
-0,12869	0,288125	0,096912	-0,22697	-0,26645	0,004523	-0,34291	-0,00579	0,153195	0,049123
-0,12869	0,288125	0,096912	-0,22697	-0,26645	0,004523	-0,34291	-0,00579	0,153195	0,049123
-0,13026	0,055378	0,052606	0,062175	0,567275	-0,43408	-0,14121	0,417507	0,496567	0,149242
-0,20646	0,491947	0,148277	0,178982	0,073682	0,082395	-0,06666	-0,05057	-0,12017	-0,05176
-0,20646	0,491947	0,148277	0,178982	0,073682	0,082395	-0,06666	-0,05057	-0,12017	-0,05176
-0,3189	-0,03884	-0,35919	0,548105	-0,36056	0,089962	0,152425	-0,14418	0,467732	0,257805
-0,10033	-0,06662	-0,00976	-0,17166	0,236607	0,865542	0,013339	0,325521	0,206191	0,046158
-0,07777	0,203823	0,051365	0,405954	0,340128	0,077872	0,276241	-0,04478	-0,27337	-0,10088
-0,51334	0,130506	-0,12317	-0,50174	-0,07732	-0,16973	0,634777	0,111535	-0,04836	0,005691
-0,32883	-0,17986	-0,18757	-0,08424	0,233636	0,012955	-0,20298	-0,422	0,217383	-0,70424
-0,52414	-0,33766	-0,16313	0,138582	-0,0625	-0,05395	-0,40118	0,357024	-0,51375	0,079226
-0,27863	-0,26157	0,419399	-0,1134	0,255004	0,049115	-0,04295	-0,56296	-0,02005	0,5278
-0,08725	-0,15008	0,428121	0,120339	-0,17718	-0,01058	0,098907	0,137639	0,096001	-0,18968
-0,08725	-0,15008	0,428121	0,120339	-0,17718	-0,01058	0,098907	0,137639	0,096001	-0,18968
-0,08725	-0,15008	0,428121	0,120339	-0,17718	-0,01058	0,098907	0,137639	0,096001	-0,18968
Σ=									
4,288913	0	0	0	0	0	0	0	0	0
0	3,425719	0	0	0	0	0	0	0	0
0	0	2,752857	0	0	0	0	0	0	0
0	0	0	2,294972	0	0	0	0	0	0
0	0	0	0	1,863937	0	0	0	0	0
0	0	0	0	0	1,346582	0	0	0	0
0	0	0	0	0	0	1,036964	0	0	0
0	0	0	0	0	0	0	0,643126	0	0
0	0	0	0	0	0	0	0	0,480506	0
0	0	0	0	0	0	0	0	0	0,131236

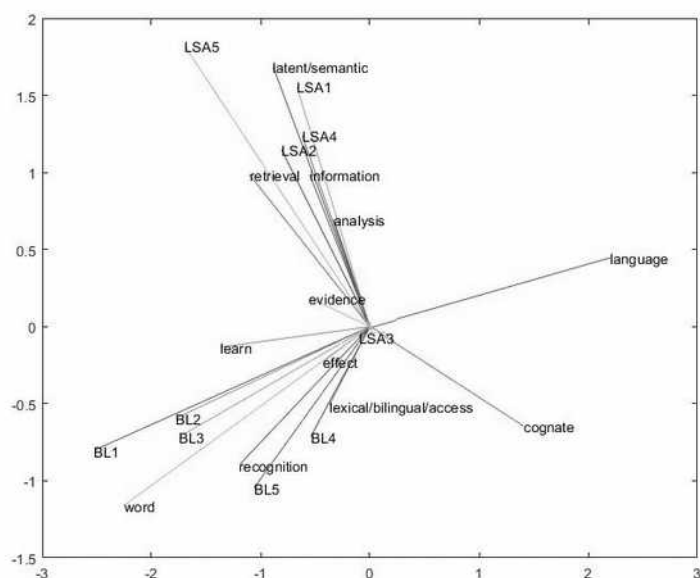
V=									
-0,15629	0,455421	0,178135	-0,04182	-0,20684	0,129094	-0,78994	-0,17529	0,137458	-0,04014
-0,18877	0,335368	-0,00409	0,571695	0,068096	0,247013	0,28481	-0,45108	-0,09568	0,406996
-0,02339	-0,01945	-0,00355	-0,0748	0,126939	0,642769	0,012864	0,506154	0,429113	0,351719
-0,14478	0,362872	0,145494	0,359958	0,565881	-0,14215	0,001642	0,422282	-0,03567	-0,42023
-0,39567	0,531613	0,088652	-0,47907	-0,2898	-0,123	0,43436	0,171564	-0,06385	0,046584
-0,58949	-0,23421	-0,49235	0,343095	-0,37008	-0,06294	-0,06338	0,179179	0,160215	-0,18658
-0,4139	-0,17316	-0,00068	-0,21727	0,491484	-0,44237	-0,14807	-0,15377	0,274276	0,439821
-0,40693	-0,20877	-0,02333	-0,31916	0,314081	0,522754	0,000969	-0,2968	-0,33004	-0,34566
-0,12599	-0,20779	0,618907	0,107894	-0,14836	0,012909	0,244724	-0,23331	0,557658	-0,31429
-0,2482	-0,30635	0,55965	0,16828	-0,18189	-0,02715	-0,14216	0,321829	-0,51153	0,2894

Ako se uzme matrica iz Tablice 2 koja prikazuje frekvencije riječi koje se pojavljuju najmanje dva puta u korpusu od deset naslova znanstvenih članaka iz Tablice 1, dobit će se tri matrice U , Σ i V koje su nastale rastavom na singularne vrijednosti kao u Tablici 3. Slijedi temeljni postupak u SVD-u, a to je smanjenje dimenzija na k . Iako smanjenje dimenzija nema ulogu da se odnose među riječima i kontekstima čini prostorno vidljivima u dvo- ili trodimenzionalnom obliku, ipak ćemo smanjiti dimenzionalnost na $k = 2$ kako bismo iz praktičnih razloga mogli zorno predočiti semantičku sličnost među riječima i kontekstima u relacijama koje čovjek može pojmiti.



Slika 2: Dvodimenzionalni prikaz rasporeda riječi i kontekstā dobiven iz rastava na singularne vrijednosti na temelju podataka iz Tablice 3. Koordinate (x, y) za riječi dobivene su množenjem stupaca matrice \mathbf{U} i $\mathbf{\Sigma}$, a koordinate za kontekste množenjem stupaca matrice \mathbf{V} i $\mathbf{\Sigma}$. Vide se dva odvojena tematska područja – dvojezično procesiranje riječi i latentna semantička analiza. Nekoliko se riječi istodobno pojavljuje u naslovima članaka. To znači da imaju iste koordinate. Stoga su prikazane jednom točkom. Tri su riječi na razmeđi između dvaju područja jer se pojavljuju u oba tematska područja. Iako se riječ evidence pojavljuje u oba tematska područja, čini se bližom području s temom o latentnoj semantičkoj analizi.

Slika 2 prikazuje prostorni razmještaj najčešćih riječi i kontekstā iz Tablice 1. Koordinate (x, y) za pojedinu riječ dobivene su tako da se za koordinatu x pomnožio element matrice \mathbf{U} iz prvoga stupca s elementom matrice $\mathbf{\Sigma}$ iz njezina prvoga stupca, a za koordinatu y pomnožio se element matrice \mathbf{U} iz drugoga stupca s elementom matrice $\mathbf{\Sigma}$ iz njezina drugoga stupca. Isti će se postupak provesti i pri određivanju koordinata za kontekste. U tom će se slučaju umjesto elemenata matrice \mathbf{U} ovaj put množiti elementi matrice \mathbf{V} s elementima matrice $\mathbf{\Sigma}$.



Slika 3: Prikaz kuta između riječi i kontekstā. Što je kut manji, to je sličnost među riječima i/ili kontekstima veća. Zbog ograničenosti korpusa proizlazi da su riječi lexical, bilingual i access kao i latent i semantic istog značenja, no iz većega bi korpusa bilo vidljivo da to ipak nisu. Posljedica je to činjenice što se te riječi pojavljuju uvijek u istom kontekstu. Potreban je stoga veći korpus da bi se razlika mogla učiniti vidljivom.

Iz Slike 2 se jasno vidi kako postoje dvije skupine koje su jasno odijeljene. Jednu skupinu čine riječi u kontekstima koje su povezane s teorijom i primjenom latentne semantičke analize, dok drugu tematsku skupinu čine riječi u kontekstima vezani uz prepoznavanje i obradu riječi kod bilingvalnih govornika. Na Slici 3 se pak vidi kako su riječi odnosno konteksti koji međusobno dijele najmanji kut najbližiji. Ovdje valja napomenuti kako riječi *lexical*, *bilingual* i *access* te *latent* i *semantic* uvijek dolaze zajedno pa ih model smatra istima odnosno isto značnicama. Da je korpus veći, ovakvo što se zasigurno ne bi dogodilo.

Prema Deerwester i sur. (1990) nastat će nova matrica A_k ako se pomnože matrice sa smanjenom dimenzionalnosti. Ta se matrica, međutim, neće u potpunosti poklapati s izvornom matricom A . No sve će se više poklapati, što se više singularnih vrijednosti ne izjednači s nulom. Ako se taj postupak primijeni na tri matrice dobivene rastavom na singularne vrijednosti, dobit će se matrica kao u Tablici 4. Ovdje se može uočiti bitna razlika u odnosu na Tablicu 3: riječi koje se nikad nisu pojavljivale u određenom kontekstu, sad to

čine. Tako se, primjerice, riječ *learn* ne pojavljuje u izvornom kontekstu BL3, no zbog toga što s ostalim kontekstima dijeli riječi poput *language*, *cognate* i *word*, izvorna vrijednost od 0 zamijenjena je vrijednošću 0.58. Ako se dalje pogleda Tablica 5 koja prikazuje vrijednosti interkorelacija kontekstā prije i poslije provođenja SVD-a, vidi se da se interkorelacija povećala kao, primjerice, između LSA1 i LSA2 s 0.318 na 0.987.

Tablica 4: Matrica A_k nastala kao umnožak triju matrica U_k , Σ_k i V_k iz Tablice 3. Tim je trima matricama smanjena dimenzionalnost sa r na k . Valja pri tome napomenuti da je matrica V_k transponirana. Tako nastala matrica A_k aproksimacija je matrice A .

$A_k=$										
	LSA1	LSA2	LSA3	LSA4	LSA5	BL1	BL2	BL3	BL4	BL5
information	0,535782	0,435211	-0,00628	0,438081	0,743113	0,094197	0,057542	0,01854	-0,13555	-0,16538
retrieval	0,535782	0,435211	-0,00628	0,438081	0,743113	0,094197	0,057542	0,01854	-0,13555	-0,16538
evidence	0,173716	0,169084	0,00938	0,149729	0,321908	0,284906	0,198393	0,187738	0,030972	0,080549
latent	0,905906	0,732341	-0,01206	0,739745	1,246282	0,127282	0,074691	0,008495	-0,23861	-0,2965
semantic	0,905906	0,732341	-0,01206	0,739745	1,246282	0,127282	0,074691	0,008495	-0,23861	-0,2965
learn	0,153168	0,213558	0,034584	0,149743	0,470436	0,837424	0,589153	0,584347	0,199972	0,380236
effect	-0,03668	0,004691	0,014505	-0,02051	0,048939	0,30712	0,217631	0,222755	0,101639	0,176722
analysis	0,370124	0,29713	-0,00578	0,301664	0,503168	0,033085	0,017149	-0,01004	-0,10306	-0,13112
language	0,547707	0,565532	0,04281	0,480994	1,108798	1,193132	0,833859	0,802573	0,184496	0,409491
cognate	-0,06019	0,059582	0,044975	-0,01939	0,230466	0,97567	0,69043	0,702531	0,305717	0,538801
word	-0,17545	0,036421	0,075083	-0,09427	0,274542	1,596081	1,130758	1,156264	0,523581	0,912318
recognition	-0,22132	-0,07494	0,045382	-0,15214	-0,00353	0,914322	0,649793	0,673365	0,336757	0,571121
lexical	-0,17567	-0,10179	0,018752	-0,13239	-0,12527	0,341	0,24391	0,259608	0,153977	0,250382
bilingual	-0,17567	-0,10179	0,018752	-0,13239	-0,12527	0,341	0,24391	0,259608	0,153977	0,250382
access	-0,17567	-0,10179	0,018752	-0,13239	-0,12527	0,341	0,24391	0,259608	0,153977	0,250382

Tablica 5: Interkorelacije (r) kontekstnih vektora između računalno neobrađenih podataka (gore) i između podataka dobivenih rekonstrukcijom matrice smanjene dimenzionalnosti (dolje). Kraticom LSA označeni su naslovi vezani uz latentnu semantičku analizu, a BL uz bilingvalnu jezičnu obradu.

	LSA1	LSA2	LSA3	LSA4	LSA5	BL1	BL2	BL3	BL4	BL5
LSA1										
LSA2	0,318182									
LSA3	-0,16116	-0,16116								
LSA4	0,318182	0,659091	-0,16116							
LSA5	0,592157	0,098693	-0,17496	0,098693						
BL1	-0,33887	0,084717	-0,15019	-0,33887	-0,061313934					
BL2	-0,4264	-0,4264	-0,18898	-0,1066	-1,28483E-17	0,39736				
BL3	-0,4264	-0,4264	0,377964	-0,4264	1,28483E-17	0,39736	0,7			
BL4	-0,36364	-0,36364	-0,16116	-0,36364	-0,394771017	-0,33887	-0,1066	-0,1066		
BL5	-0,4264	-0,4264	-0,18898	-0,4264	-0,46291005	0	0,1	0,1	0,852803	

	LSA1	LSA2	LSA3	LSA4	LSA5	BL1	BL2	BL3	BL4	BL5
LSA1										
LSA2	0,987309711									
LSA3	-0,65829404	-0,5303967								
LSA4	0,99916455	0,992974989	-0,60425992							
LSA5	0,95955616	0,992085915	-0,4197549	0,970259579						
BL1	-0,42475711	-0,27559804	0,96109504	-0,38740404	-0,15271865					
BL2	-0,44093814	-0,29280763	0,965898071	-0,40388909	-0,17043259	0,9998389				
BL3	-0,50642871	-0,36306594	0,982471041	-0,47076578	-0,24319956	0,99554361	0,997234			
BL4	-0,81927591	-0,71781944	0,970957163	-0,79515768	-0,62471938	0,86709605	0,875898	0,90934		
BL5	-0,76151891	-0,64892568	0,989199827	-0,73439433	-0,5482568	0,91022892	0,917515	0,94454	0,99554	

6.2. Hiperprostor analogan jeziku (HAL)

HAL je drugi poznati distribucijski semantički model na kojem su uglavnom radili Kevin Lund i Curt Burgess od 1992. Model HAL ima mnogo poveznica s prethodnim modelom LSA-om, no ima i neke značajne razlike. Oba modela iz tekstova pokušavaju izlučiti bitne informacije iz matrica koje su sazdane od riječi koje se usporedno pojavljuju. Postoje dvije ključne razlike između ta dva modela:

1. Za razliku od LSA-a u kojem su elementi matrice usporednoga pojavljivanja jedinice koje čine presjek riječi i multiskupa (odlomak, dio teksta, cijeli tekst ili sve ono što pojedini istraživač smatra svrsishodnim u svojem radu), u HAL-u su ti elementi presjeci usporednoga pojavljivanja dviju riječi. Drugim riječima, ova dva modela imaju različito shvaćanje konteksta.
2. Dok su kod LSA-a sve riječi unutar istoga konteksta jednako važne, kod HAL-a svaka riječ u kontekstu ima drugačije postavljen težinski faktor tako da nije jednako važna (Burgess, Livesay i Lund, 1998).

Lund, Burgess i Atchley (1995) ističu da je model HAL nastao pod utjecajem Schützeova rada (1992). Upravo tu tvrdnju osporavaju Landauer i Dumais (1997) ističući kako Schütze u tom i u drugim svojim člancima citira rad Deerwestera i sur. (1990) na kojemu počiva ideja LSA-a. HAL je takorekuć izdanak LSA-a.

HAL se ne oslanja na ljudske odluke u konstrukciji matrice niti to čini na nadziran način (*unsupervised fashion*) kao što to ne čini ni LSA. Matrice su višedimenzionalne i konstruirane od nekoliko milijuna pojavnica. Često se takve višedimenzionalne prostore naziva kontekstnim prostorima (Burgess, Livesay i Lund, 1998; Burgess i Lund, 1998; Burgess, 1998). Svaki element matrice predstavlja simbol. Uobičajeno je to riječ, no mogu biti i brojke, emotikoni i slično ovisno o tome kakva je vrsta teksta (Burgess, 1998). Nastala će matrica kao i kod LSA-a biti vrlo rijetka (*sparse*) što znači da će njezini elementi uglavnom biti nule.

Da bi HAL izradio matricu, koristi se pokretnim prozorom (*moving window*) koji prolazi kroz korpus zahvaćajući unaprijed određen broj riječi koje se nalaze ispred i iza središnje riječi. Takvi se spojevi više uzastopnih riječi nazivaju n-gramima. Prozor Burgess i Lund (1997) određuju kao najveći broj riječi koji se pojavljuju između dva para riječi. Riječi koje se pojavljuju unutar prozora smatraju se lokalnim pojavnostima koje se međusobno mogu usporediti. Najmanji prozor može biti jedan, što znači da obuhvaća riječ koja

neposredno prethodi i slijedi središnju riječ.¹ Gornja granica veličine prozora nije određena. Veličina je prozora u nekoliko navrata bila predmet istraživanja (Li, Burgess i Lund, 2000; Lund i Burgess, 1996). Ako se koristi vrlo malenim prozorima, postoji opasnost da se propuste konstrukcije od više riječi poput nekih fraza i slično. Isto se tako korištenjem velikih prozora zahvaćaju nepotrebne riječi. Stoga su se Burgess i Lund u svojim istraživanjima uglavnom koristili veličinom prozora od deset riječi (Burgess, 1998; Burgess, Livesay i Lund, 1998; Burgess i Lund, 1998; Li, Burgess i Lund, 2000).

Rečenična se interpunkcija ne uzima u obzir. Kako pokretni prozor prolazi kroz rečenicu, zanemaruje točku koja označava kraj rečenice te obuhvaća i riječi sljedeće rečenice.

Pomičući pokretni prozor od riječi do riječi, svakoj se riječi određuje težinski faktor (*weighting*) ovisno o tome koliko je blizu središnjoj riječi. Ako se uzme prozor veličine pet, to znači da riječ koja neposredno prethodi središnjoj riječi i neposredno slijedi središnju riječ dobiva težinski faktor pet. Zadnja pak riječ u prozoru dobiva težinski faktor jedan. Ako se u istom prozoru dva puta pojavi ista riječ, vrijednosti se težinskih faktora zbrajaju. Redci matrice predstavljaju tako riječi koje su prethodile središnjoj riječi, a stupci predstavljaju riječi koje ju slijede.

Kako bismo zorno prikazali kako HAL konstruira matricu, uzet ćemo kao primjer prvu rečenicu romana Danijela Bogdanovića "Noćni vlak za Dukku" (2015: 11): "U mom snu ja vidim velike strojeve, strahovite strojeve, jezovite, s puno dugačkih nogu koje se miču i vitlaju po zraku i mlate i strahovite su mi." Tablica 6 prikazuje kvadratičnu matricu tipa (25×25) usporednoga pojavljivanja riječi s pokretnim prozor veličine pet. U ovoj se tablici nekoliko riječi pojavljuje više puta. Ako se kao središnja riječ uzme riječ *jezovite*, tada se na zarez postavlja težinski faktor pet², na riječ *strojeve* težinski faktor četiri i tako redom sve do zadnje riječi u ovom prozoru koju ponovno čini riječ *strojeve*, a ona se vrednuje s jedan. Težinski se faktori za riječ *strojeve* zbrajaju, pa će konačna vrijednost iznositi pet.

¹ Kod samoga određivanja najmanje veličine pokretnoga prozora Burgess i Lund prilično su nekonzistentni. Primjerice, u Burgess i Lund (1997) ističu da je donja granica prozora jedan. To znači da jedna riječ prethodi središnjoj riječi prozora i jedna ju slijedi. U Burgess, Livesay i Lund (1998) najmanja se veličina prozora definira kao dva uzimajući u obzir riječ koja neposredno prethodi i slijedi središnju riječ.

² Premda se interpunkcijski znakovi ne uzimaju u obzir, zbog malenoga konteksta od samo jedne rečenice, a i radi zornosti, ipak ćemo ih uzeti u obzir.

Tablica 6: Matrica prema modelu HAL. Kao egemplaran je korpus uzeta prva rečenica romana "Noćni vlak za Dukku" Danijela Bogdanovića (2015: 11): "U mom snu ja vidim velike strojeve, strahovite strojeve, jezovite, s puno dugačkih nogu koje se miču i vitlaju po zraku i mlate i strahovite su mi". Prikazana je matrica dobivena pokretnim prozorom veličine 5. Redci označavaju vrijednosti usporednoga pojavljivanja riječi koje prethode riječi na koju se redak odnosi, a stupci označavaju vrijednosti usporednoga pojavljivanja riječi koje slijede riječ na koju se stupac odnosi.

	u	mom	snu	ja	vidim	velike	strojeve	strahovite	jezovite	s	puno	dugačkih	nogu	koje	se	miču	i	vitlaju	po	zraku	mlate	su	mi	,	.
u	0	5	4	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mom	0	0	5	4	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
snu	0	0	0	5	4	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
ja	0	0	0	0	5	4	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
vidim	0	0	0	0	0	5	5	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
velike	0	0	0	0	0	0	7	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
strojeve	0	0	0	0	0	0	3	4	5	2	1	0	0	0	0	0	0	0	0	0	0	0	0	15	0
strahovite	0	0	0	0	0	0	5	0	3	1	0	0	0	0	0	0	0	0	0	0	0	5	4	6	3
jezovite	0	0	0	0	0	0	0	0	0	4	3	2	1	0	0	0	0	0	0	0	0	0	0	5	0
s	0	0	0	0	0	0	0	0	0	0	5	4	3	2	1	0	0	0	0	0	0	0	0	0	0
puno	0	0	0	0	0	0	0	0	0	0	0	5	4	3	2	1	0	0	0	0	0	0	0	0	0
dugačkih	0	0	0	0	0	0	0	0	0	0	0	0	5	4	3	2	1	0	0	0	0	0	0	0	0
nogu	0	0	0	0	0	0	0	0	0	0	0	0	0	5	4	3	2	1	0	0	0	0	0	0	0
koje	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	4	3	2	1	0	0	0	0	0	0
se	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	4	3	2	1	0	0	0	0	0
miču	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	4	3	2	0	0	0	0	0
i	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	6	5	4	3	6	6	4	0	2
vitlaju	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	5	4	2	0	0	0	0

Matrica koja nastaje na ovakav način zapravo je spoj dviju trokutastih matrica. Svaki je element matrice zbroj vrijednosti usporednoga pojavljivanja dviju riječi. Ako se iz Tablice 6 uzmu tako dvije riječi *strojeve* i *strahovite*, skup (*strojeve*, *strahovite*) i skup (*strahovite*, *strojeve*) neće imati istu vrijednost.

U ovakvoj matrici tipa ($N \times N$) N predstavlja broj riječi koji se uzeo u obzir. Puni vektor usporednoga pojavljivanja neke riječi čini ulančanje elemenata matrice retka i stupca. Tako bi u Tablici 6 vektor riječi *strahovite* izgledao ovako:

[0 0 0 0 0 0 5 0 3 1 0 0 0 0 0 0 0 0 0 0 0 0 5 4 6 3 0 0 0 1 2 3 4 0 0 0 0 0 0 0 0 0 8 0 1 2 4 0 0 5 0]

Ti se vektori mogu promatrati kao koordinate u višedimenzionalnom prostoru gdje svaka riječ zauzima svoje mjesto. Mjerenjem udaljenosti između dvaju vektora može se odrediti semantička razlika između tih dviju riječi. Kako bi se to moglo učiniti, vektore prvo treba normalizirati. To znači da valja prvo odrediti duljinu ili normu vektora $\|a\|$ koja predstavlja zbroj kvadrata njegovih elemenata:

$$\|a\| = \sqrt{\sum_{i=1}^n a_i^2}$$

Vektor se a potom dijeli s normom $\|a\|$ kako bi se dobio unitarni vektor \hat{a} :

$$\hat{a} = \frac{a}{\|a\|}$$

Na taj se način uklanja sva moguća prednost koju imaju vektori riječi koji se češće pojavljuju. Svi unitarni vektori imaju istu duljinu. Jedino što ih razlikuje jest smjer (Widdows, 2004). Unitarni se vektor \hat{a} dalje dijeli sa 666.0. Takva se mjera udaljenosti naziva *Riverside context unit* (Burgess, 1998; Burgess, Livesay i Lund, 1998).

Udaljenost između dvaju vektora mjeri se minkovskijevom udaljenosti koja je osjetljiva na magnitudu vektora zbog čega ih se prvo mora normalizirati. Minkovskijeva se udaljenost može smatrati generalizacijom euklidske udaljenosti i l_1 -metričke funkcije ili Manhattan-metrike, a prikazuje se sljedećom formulom:

$$d(x, y) = \sqrt[r]{\sum_{i=1}^n (|x_i - y_i|)^r}$$

Pri tomu je $r = 2$ ako se radi o euklidskoj udaljenosti, a $r = 1$ ako se radi o l_1 -metričkoj funkciji (Lund i Burgess, 1996). U trećem eksperimentu opisanom u Lund i Burgess (1996) ispitivali su koja je optimalna veličina pokretnoga prozora te optimalan iznos r . Najbolji su rezultati postignuti s veličinom prozora od 8 i r od 1.

Budući da se u Tablici 6 radi o relativno preglednoj matrici s definiranim brojem riječi, a svaku riječ predstavlja vektor koji sadrži sveukupno 50 elemenata, nužno je postaviti neka ograničenja. Naime, Burgess i Lund u više su svojih istraživanja zaključili da za simulaciju ljudskoga uma nije potrebno nekoliko stotina tisuća dimenzija, kako je u distribucijskim modelima često slučaj, već je dovoljno 100-200 dimenzija (Burgess, Livesay i Lund, 1998; Burgess i Lund, 1997; Lund, Burgess i Atchley, 1995). Burgess (1998) tako predlaže skraćivanje vektora na 200 dimenzija ne bi li se tako zadržali oni elementi vektora koji su "najinformativniji". Da bi se to moglo odrediti, treba se izračunati varijancu između stupca i retka. Elementi se s najmanjom varijancom isključuju jer predstavljaju kontekst u kojem se dotična riječ vrlo rijetko pojavljuje, tj. kontekstno joj je iskustvo vrlo skromno. Burgess nadalje ističe da je varijanca kao mjera informativnosti pogodna za HAL jer je odraz različitosti konteksta u kojima se pojavljuju riječi. Međutim, u svojim istraživanjima Burgess se i dalje koristi neskraćenim vektorima. Pokazalo se da je to računalno učinkovitije jer se izbjegava računanje varijance za velike matrice.

Sa svakim novim pojavljivanjem pojedine riječi u kontekstu povećava se njezino kontekstno iskustvo. Na taj se način obogaćuje kontekstni vektor (Burgess i Lund, 1998). To znači da se svaki put povećava broj dimenzija kada se u novom kontekstu otkrije određena riječ. Ako se poslužimo Tablicom 6, ukoliko se bilo koja riječ iz te rečenice pojavi u nekom novom kontekstu, tim se događajem ne povećava samo broj stupaca i redaka u matrici već se povećavaju i pozitivni elementi vektora. Iz niza lokalnih usporednih pojavljivanja nastaje na taj način pojava globalnoga usporednog pojavljivanja. Drugim riječima, pojava je globalnoga usporednog pojavljivanja ulančanje svih pojava lokalnoga usporednog pojavljivanja. Globalno je usporedno pojavljivanje, dakle, ukupno kontekstno iskustvo pojedine riječi koje je ta riječ stekla usporednim pojavljivanjem s drugim riječima unutar raspona pokretnoga prozora (Burgess, Livesay i Lund, 1998). Burgess (1998) pojavu globalnoga usporednog pojavljivanja naziva i "kontekstnom poviješću riječi".

Kako bi se prostorno prikazali vektori, koristi se tehnikom višedimenzionalnoga skaliranja (*multidimensional scaling algorithm*, MDS). Tom se tehnikom vektore projicira iz

višedimenzionalnoga prostora na niskodimenzionalni prostor, a da se pri tome zadržavaju svi prostorni odnosi (Burgess i Lund, 1998). Na taj se način pokušavaju zorno prikazati odnosi među pojedinim riječima u dvodimenzionalnom formatu.

Bullinaria i Levy (2007, 2012) otkrili su da je za DSM-ove koji broje pojavnost riječi poput HAL-a bitna veličina prozora. Ovisno o zadatku razlikuje se optimalna veličina prozora. To je, primjerice, za semantičko usmjeravanje 3, za zadatak donošenja odluka o semantičkoj povezanosti 1, a za predviđanje asocijacija 2.

6.3. Vezano enkodiranje cjelokupne jezične okoline (BEAGLE)

Jedan je od najčešćih prigovora DSM-u bio taj da nije utemeljen u izvanjezičnoj stvarnosti, a LSA i HAL uz to ne uvažavaju sintaksu koja je od ključne važnosti u ljudskoj jezičnoj obradi. Zbog toga su Wiemer-Hastings i Zipitria (2001) sa svojim modelom strukturiranoga LSA-a (*structured LSA – SLSA*) te Serafin i Di Eugenio (2004) sa svojim modelom koji uzima u obzir semantička obilježja u LSA-u (*feature LSA – FLSA*) pokušali riješiti taj problem. Wiemer-Hastings i Zipitria su tako složene rečenice rastavljali na jednostavnije, dodjeljivanjem rečeničnih uloga (npr. subjekt, predikat, objekt) pojedinim riječima i skupovima riječi ili uklanjanjem pronominalne anafore, zamijenivši zamjenicu riječju na koju se odnosi. Serafin i Di Eugenio su, s druge strane, primijenili svoj model FLSA na klasifikaciju dijaloških činova. Na taj se način mogu razlikovati uloge koje igraju pojedini sudionici u dijalogu.

Perfetti (1998) je već netom nakon objavljivanja temeljnih članaka koje čine okosnicu modela LSA-a i HAL-a zaključio da ni jedan ni drugi ne vode računa o sintaksi. U svojemu članku tvrdi osim toga da ne odražavaju ljudsku kogniciju. Prozor veličine od deset riječi od središnje riječi iz modela HAL ne odražava radnu memoriju kod ljudi kako to tvrde Burgess, Livesay i Lund (1998). Perfetti to potkrepljuje istraživanjima Goldman i sur. (1980) koji su upravo to ustanovili. Nadalje tvrdi da model HAL ne utjelovljuje ljudsko semantičko znanje na što upućuju provedeni eksperimenti, iako hipoteze na kojima počiva model imaju sintaktičke i leksičko-sintaktičke karakteristike. Također pretpostavka je da LSA radi s napisanim i pregledanim tekstovima što znači da sintaksa nije bitna jer su tekstovi već oblikovani prema sintaktičkim pravilima pojedinoga jezika.

Kako bi riješili taj problem, Jones, Kintsch i Mewhort (2006) predlažu drugačiji model koji spaja modele LSA i HAL. Informacije koje prikupljaju oba modela skladište se u

kompozitnom holografskom leksikonu. Kontekst su diskretne rečenice, odnosno cjelovite sintaktičke strukture. Svoj su model nazvali BEAGLE što je skraćeno za *Bound Encoding of the Aggregate Language Environment* ili *Vezano enkodiranje cjelokupne jezične okoline*. BEAGLE se temelji na Murdockovu radu (1982) koji je upotrijebio konvoluciju za upareno asocijativno učenje. Konvolucija je metoda komprimiranja matrice koja je nastala vektorskim umnoškom dvaju vektora. Inverzna operacija konvoluciji je korelacija ili dekonvolucija. Korelacijom se iz nastale matrice mogu dekodirati njezine približne izvorne sastavnice. Takvi se konvolucijsko-korelacijski modeli često nazivaju holografskim modelima (Plate, 2003). Murdock (1982) operacije konvolucije i korelacije opisuje vrlo slikovito. Tako, primjerice, konvoluciju uspoređuje s valovima koje na površini jezera uzrokuju razni predmeti koji se bace u njega. Ako se u jezero, naime, bace automobilski kotač, boca piva, sudoper i telefon, na površini će jezera nastati različiti valovi. Međutim, ti valovi neće biti specifični ni za jedan predmet. Uspoređivanjem se pak može zaključiti da je u jezero ubačen telefon, a nisu skije. Isto bi tako bilo moguće "dohvatiti informaciju" da je ubačena boca piva ako kao mig posluži kotač. Murdock ističe da se ne pretražuju sami predmeti, već informacije koje su dostatne za vjerodostojno prepoznavanje. Pohranjivanje informacija bi tako bila konvolucija, a pretraživanje informacija korelacija.

Memorijski modeli uglavnom rabe konvoluciju kako bi kodirali poredak riječi u jeziku koristeći se ogromnim statističkim podacima prikupljenim iz korpusa. Budući da se često koristi aperiodičnom konvolucijom, dijagonale se matrice vektorskoga umnoška zbrajaju. Nastali se rezultat ne može usporediti s izvornim elementima jer je veće dimenzionalnosti. To je veliki problem jer se dimenzionalnost vrlo brzo povećava rekurzivnom aperiodičnom konvolucijom. Kako bi se, naime, asocijiralo riječi unutar rečenice, uzastopni se n -grami (spojevi od dva, tri, četiri i više uzastopnih riječi) moraju međusobno rekurzivno povezati. Da bi se riješio problem rastuće dimenzionalnosti rekurzivnom konvolucijom, BEAGLE se koristi cirkularnom konvolucijom. Primjenom takve tehnike dva se vektora asocijiraju u jedan vektor jednake dimenzionalnosti koji će poslužiti kao izvorišni (Jones i Mewhort, 2007).

Cirkularnom konvolucijom nastaju jedinstveni asocijativni vektori. Tako se, primjerice, asocijativni vektori $a \cdot b$ i $a \cdot b \cdot c$ nastali od bigrama odnosno trigrama međusobno mogu usporediti jer su jednake dimenzionalnosti. Međutim, n -grami višega reda sadržavaju drugačije informacije od onih nižega reda. N -grami se višega reda, naime, češće pojavljuju u takvoj kombinaciji. Zbog toga su n -grami nižega reda važniji za povijest poretka riječi.

Činjenica kako je konvolucija komutativna ($a \cdot b = b \cdot a$) predstavlja daljnji problem. Često je, naime, tako da te dvije opcije nemaju jednako značenje. Da bi se riješio taj problem, konvolucija se mora učiniti nekomutativnom ($a \cdot b \neq b \cdot a$). Plate (1995) predlaže da se prije nego što se dva vektora konvoluiraju, njihove elemente različito permutira. Na taj će način $a \cdot b$ i $b \cdot a$ biti svaki za sebe jedinstveni asocijativni vektori.

U nastavku će se opisati osnovne značajke modela BEAGLE. Pri opisu ćemo se uglavnom voditi dvama radovima: Jones i Mewhort (2007) te Jones, Kintsch i Mewhort (2006). U modelu BEAGLE polazi se od toga da su značenje riječi i uporaba riječi jedinstven obrazac asociran u jednom vektoru. U početku je svakom vektoru dodijeljena nasumična vrijednost. Svakim sljedećim pojavljivanjem riječi u rečenici vektor te riječi dobiva na kontekstualnom iskustvu i na taj se način smanjuju smetnje (*noise*), odnosno uvodi se red u početnu kaotičnu strukturu.

Riječ je višedimenzionalni holografski vektor. Prilikom otkrivanja nove riječi prvo joj se pripisuje nasumičan okružni vektor e_i (*environmental vector*). Okružni su vektori odabrani nasumično prema Gaussovoj ili normalnoj distribuciji. Jones, Kintsch i Mewhort (2006) napominju da se dimenzionalnost može odrediti proizvoljno. Međutim, mora postojati dovoljan broj dimenzija kako bi se primjereno mogle pohraniti informacije o značenju riječi i o poretku riječi. Ističu da povećanje dimenzionalnosti neće niti poboljšati niti pogoršati odnose sličnosti među riječima unutar leksikona. Odnosi sličnosti među riječima jednaki su neovisno o tome je li model utreniran s dimenzionalnošću od 1000 ili od 10.000. Povećanjem se dimenzionalnosti informacije samo raspoređuju na više elemenata. Jones, Kintsch i Mewhort (2006) uspoređuju to s konektivističkim modelima koji zahtijevaju određen broj čvorova. Povećanjem broja čvorova ne smanjuje se učinkovitost. U LSA-u se za razliku od BEAGLE-a povećanjem dimenzionalnosti iznad 300 smanjila učinkovitost modela (Landauer i Dumais, 1997). Stoga Jones, Kintsch i Mewhort (2006) preporučuju da donja granica dimenzionalnosti bude 1000. U svojim su istraživanjima rabili 1024 ili 2048 dimenzija.

Svaki sljedeći put kada se model susretne s pojedinom riječi isti se okružni vektor koristi kako bi se kodirali podaci za kontekst i poredak riječi. Okružni se vektor pri tome ne mijenja. Ono što se mijenja jest memorijski odnosno leksički vektor riječi m_i , i to svaki puta kada se riječ otkrije u novom kontekstu. Okružni vektor tako predstavlja fizičke karakteristike riječi u svojem okruženju (ortografija, fonologija i dr.), dok memorijski vektor predstavlja unutarnju memoriju riječi, odnosno kontekst i položaj riječi u kojem se nalazi okružni vektor.

Svaki vektor u modelu BEAGLE kodira informacije o kontekstu (c_i) i informacije o poretku riječi (o_i) u isti memorijski vektor:

$$m_i = m_i + c_i + o_i$$

Informacije o kontekstu (c_i) u nekoj rečenici jesu zbroj okružnih vektora za ostalih $n - 1$ riječi u rečenici:

$$c_i = \sum_{j=1}^n e_j, i \neq j$$

Za svaku se riječ u rečenici računa kontekstni vektor. Tim se podatkom onda aktualizira memorijski vektor riječi pribrajući mu ga (prvi dio prethodne jednačbe: $m_i = m_i + c_i$).

Visokofrekventne riječi mogu negativno utjecati na leksički prikaz, tj. mogu biti susjedne riječi gotovo svakoj drugoj riječi u leksikonu. Da bi se to izbjeglo, upotrebljava se diferencijalna entropija. Na taj se način smanjuje utjecaj visokofrekventnih riječi čim model primijeti da se pojavljuju iznimno često. Model bi sâm trebao biti u stanju da raspoznaje riječi koje pridonose semantičkoj informaciji rečenice, a koje ne, pod uvjetom da raspolaže s dovoljno velikim brojem uzoraka rečenica odnosno dovoljno velikim korpusom. Međutim, kako bi se uštedjelo na vremenu i kapacitetu računala te ubrzala računalna obrada podataka, pribjeglo se upotrebi popisu zaustavnih riječi (*stop words*).

Ako se, primjerice, uzme prva rečenica romana "Sat njemačkoga" Siegfrieda Lenza – "Zadali su mi kaznenu zadaću" (Lenz, 1999: 9) –, kontekstni bi vektor za riječ *zadali* bio zbroj ostalih vektora kada se zanemare pomoćni glagol *su* i zamjenica *mi* koji su na popisu zaustavnih riječi:

$$c_{zadali} = e_{kaznenu} + e_{zadaću}$$

Kontekst riječi *zadali* je superpozicija riječi *kaznenu* i *zadaću*. Ta se informacija o kontekstu potom pribraja memorijskom vektoru za riječ *zadali* (m_{zadali}).

Jonas i Mewhort (2007) razlikuju četiri zavisna izvora statističkih informacija. To su informacije o usporednom pojavljivanju i informacije o tranziciji. Obje vrste informacija mogu biti izravne i neizravne. Informacije o usporednom pojavljivanju obuhvaćaju podatke o drugim riječima koje se pojavljuju unutar istoga konteksta (rečenice), a informacije o

tranziciji obuhvaćaju podatke o položaju riječi u odnosu prema drugim riječima u kontekstu. Drugim riječima, informacije o izravnom usporednom pojavljivanju i tranziciji mogle bi se nazvati sintagmatskim odnosima, a informacije o neizravnom usporednom pojavljivanju i tranziciji paradigmatičkim odnosima.

Primijenjeno na Lenzovu rečenicu, to znači da je vektor m_{zadali} izravno povezan s dvama ostalim vektorima e_{kaznenu} i $e_{\text{zadaću}}$. U vektoru m_{zadali} zapisane su, dakle, informacije o izravnom usporednom pojavljivanju s ostala dva vektora. Međutim, vektor m_{zadali} ima i neizravne informacije o usporednom pojavljivanju kada se, primjerice, pojavi s vektorima e_{brige} , e_{problemi} , e_{udarac} , e_{naredbu} i sl. Važno je napomenuti da se kontekstnim vektorima obuhvaćaju semantički odnosi.

Informacije o poretku riječi (o_i) računaju se tako da se povezuju svi n -grami u rečenicama u kojima se pojavljuje riječ. To se postiže cirkularnom konvolucijom. Položaj se riječi za koju se računa informacija o poretku riječi označava nasumičnim konstantnim vektorom koji čuva mjesto toj riječi unutar rečenice. Tom se metodom kodiraju asocijacije za n -grame koje čine susjedstvo riječi. Konvolucije su n -grama jedinstvene iako često zadržavaju zajedničke elemente. Drugim riječima, informacija o poretku riječi u rečenici zbroj je svih n -gram konvolucija:

$$o_i = \sum_{j=1}^{p\lambda - (p^2 - p) - 1} \text{bind}_{i,j}$$

gdje je p položaj riječi u rečenici, $\text{bind}_{i,j}$ je j -ta konvolucija riječi koja se kodira, a λ je parametar povezivanja kojim se može odrediti maksimalan broj susjeda s kojima će se povezati riječ koja se kodira. Budući da se broj konvolucija jako povećava u dugačkim rečenicama, obrada svih mogućih konvolucija zahtijeva mnogo vremena za računalnu obradu. Jones i Mewhort (2007) su se u svojim istraživanjima odlučili da λ ima vrijednost sedam vodeći se istraživanjem Millera i Selfridge (1949). Oni su u svojem istraživanju provjeravali koliko elemenata pamte ispitanici u nizu riječi koji je smisljeno povezan. Pokazalo se da najbolje rezultate polučuje skup od sedam riječi koji čini smislenu cjelinu.

Ako se želi kodirati informacija o poretku riječi za riječ *kaznenu* iz Lenzove rečenice "Zadali su mi kaznenu zadaću" (Lenz, 1999: 9), za razliku od određivanja kontekstnoga

vektora ne smiju se ispustiti zaustavne riječi jer su bitne u sintaktičkoj strukturi. Na Slici 4 vidi se kako izgleda konvolucija za sve n -grame koji okružuju riječ *kaznenu*.

$$\begin{aligned}
 & \left. \begin{aligned} bind_{kaznenu,1} &= e_{mi} \otimes \Phi \\ bind_{kaznenu,2} &= \Phi \otimes e_{zadacu} \end{aligned} \right\} \text{bigram} \\
 & \left. \begin{aligned} bind_{kaznenu,3} &= e_{mi} \otimes \Phi \otimes e_{zadacu} \\ bind_{kaznenu,4} &= e_{su} \otimes e_{mi} \otimes \Phi \end{aligned} \right\} \text{trigram} \\
 & \left. \begin{aligned} bind_{kaznenu,5} &= e_{su} \otimes e_{mi} \otimes \Phi \otimes e_{zadacu} \\ bind_{kaznenu,6} &= e_{zadali} \otimes e_{su} \otimes e_{mi} \otimes \Phi \end{aligned} \right\} \text{kvadrigram} \\
 & \left. \begin{aligned} bind_{kaznenu,7} &= e_{zadali} \otimes e_{su} \otimes e_{mi} \otimes \Phi \otimes e_{zadacu} \end{aligned} \right\} \text{tetragram} \\
 & o_{kaznenu} = \sum_{j=1}^7 bind_{kaznenu,j}
 \end{aligned}$$

Slika 4: Prikaz konvolucije za sve n -grame koji okružuju riječ *kaznenu* iz rečenice "Zadali su mi kaznenu zadacu". To je prva rečenica iz Lenzova romana "Sat njemačkoga" (Lenz, 1999: 9). Prilagođeno prema Jonesu i Mewhortu (2007).

U vektoru poretka riječi kodirane su informacije o izravnoj i neizravnoj tranziciji. To znači da na Slici 4 riječ *kaznenu* ima izravne tranzicijske odnose s riječi *mi* koja joj prethodi i s riječi *zadacu* koja ju slijedi. Međutim, može imati i neizravne odnose. Tako je riječ *kaznenu* slična riječi *domaću*, *školsku* i sl. Sve te riječi mogu okružiti *mi* i *zadacu*. Bitno je istaknuti da se tranzicijskim informacijama može odrediti vrsta riječi i sintaktička uloga.

Tako se dobivena informacija o poretku riječi pribraja memorijskom i kontekstnom vektoru ($m_i = m_i + c_i + o_i$):

$$m_i = m_i + \sum_{j=1}^n e_j + \sum_{j=1}^{p\lambda - (p^2 - p) - 1} bind_{i,j}$$

Valja napomenuti da su kontekstni vektor i vektor o poretku riječi normalizirani kako jedan od njih ne bi nadjačao drugoga zbog toga što bi imali različite duljine ovisno o duljini rečenice.

Dok LSA i BEAGLE imaju guste distribucijske reprezentacije (*dense distributed representations*), HAL ima rijetke (*sparse distributed representations*). Isto je tako mjera za

sličnost u LSA-u i BEAGLE-u kosinus kuta između dvaju vektora, dok je to u HAL-u jedna vrsta euklidske udaljenosti. U slučaju LSA-a i BEAGLE-a to znači što je kut između dviju riječi manji, to je vrijednost kosinusa veća, a istodobno će i sličnost među njima biti veća. Za razliku od toga kod HAL-a je sličnost veća što je udaljenost između dvaju vektora manja (Jones, Kintsch i Mewhort, 2006).

BEAGLE se mora koristiti rečenicom kao kontekstom jer se samo unutar rečenice mogu kalkulirati informacije o poretku riječi da bi se mogle odrediti točne sintaktičke veze. Drugim riječima, algoritam za kalkuliranje informacija o poretku riječi naučio je informacije o vrstama riječi. Za razliku od njega algoritam za kalkuliranje konteksta naučio je informacije o semantičkim vezama među riječima. I sâm Landauer, kako Jones i Mewhort (2007) prenose, tvrdi da bi se modelu LSA mogao dodati nasumični kontekstni vektor koji bi predstavljao odlomak kako bi se kalkulirao rečenični kontekst za svaku riječ u rečenici. Na taj bi način riječi unutar odlomka postale sličnije.

Matrica je u LSA-u i HAL-u nepromijenjena tijekom više serija računalne obrade, dok se kod BEAGLE-a mijenja. To znači da su leksikoni koje je oblikovao BEAGLE različiti od serije do serije, no struktura odnosa sličnosti ostaje neočekivano slična (Jones i Mewhort, 2007).

Dok su kod LSA-a i HAL-a višeznačne riječi predstavljale problem jer modeli nisu dobro mogli odrediti na koje se značenje misli u konkretnom slučaju, BEAGLE je u svojim vektorima kodirao sva značenja pojedine riječi zajedno s kontekstima. Kontekst u kojem se riječ upotrebljava otkrit će o kojem se značenju višeznačne riječi zapravo radi (Jones i Mewhort, 2007). To odražava i način kako ljudi postupaju u takvim situacijama. Pokazalo se, naime, da se i kod ljudi aktiviraju sva značenja višeznačne riječi, te se zatim tijekom procesa obrade određuje traženo značenje. Slično je kod višejezičnih govornika kod kojih se aktivira značenje određene riječi i u drugim jezicima koje poznaje (Onifer i Swinney, 1981; Titone i sur., 2011).

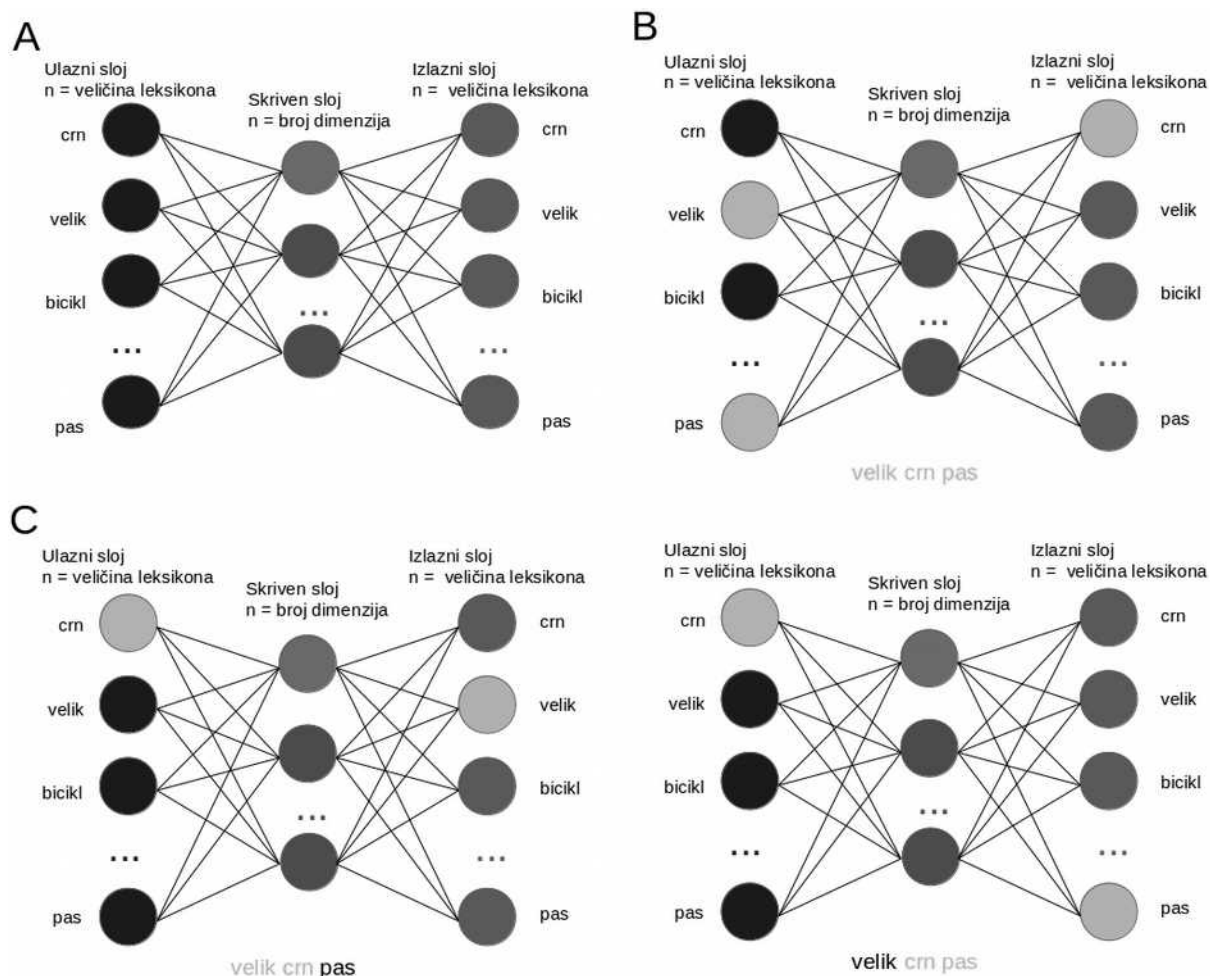
6.4. *Word2vec*

Kada se govori o modelu *word2vec*, često se misli da je to jedinstven algoritam poput prije spomenutih LSA-a ili HAL-a. *Word2vec* nije jedinstven monolitan algoritam. Sastoji se od dva odvojena modela koji se opet, svaki za sebe, razlikuju po tome koji su parametri odabrani ili koja je varijacija odabrana. Ta su dva modela kontinuirani skip-gram model i

model kontinuiranoga multiskupa riječi (*Continuous Bag-of-Words model*, CBOW) (vidi sliku 5). Oba modela su kriptično izložena u Mikolov, Chen, Corrado i Dean (2013a) i u Mikolov, Sutskever, Chen, Corrado i Dean (2013b). Druga je razlika prema prijašnjim modelima LSA-u, HAL-u i BEAGLE-u ta što su to modeli koji se temelje na prebrojavanju (*count models*), dok je *word2vec* model koji se temelji na predviđanju (*predictive models*) (Baroni, Dinu i Kruszewski, 2014). Jednu sličnost, međutim, dijeli *word2vec* s modelom HAL, a ta je što se oba modela koriste prozorom stanovite veličine oko središnje riječi koji se pomiče kroz korpus tekstova (Mandera, Keuleers i Brysbaert, 2017).

Word2vec nije duboko učenje (*deep learning*), već je "plitak" neuronski model. Drugim riječima, distribuirane su reprezentacije riječi nastale na temelju učenja umjetnih neuronskih mreža (Mikolov i sur., 2013a). Pokazalo se da takvi modeli daju bolje rezultate od drugih DSM-ova poput LSA-a u zadacima gdje model treba odrediti oblik množine ako je zadan oblik jednine, oblike glagola u prošlosti ako je zadan oblik sadašnjosti i sl. (Mikolov, Yih i Zweig, 2013).

Cilj je Mikolova i sur. (2013a) bio taj da nađu takav model koji se temelji na neuronskim mrežama poput unaprijednih neuronskih mreža (*feedforward neural net language model*) ili rekurentne neuronske mreže (*recurrent neural net language model*), ali da bude mnogo jednostavniji. Takav bi model trebao predstavljati podatke podjednako precizno kao modeli temeljeni na neuronskim mrežama, no bio bi istodobno mnogo učinkovitiji. Otkrili su da se jezični modeli temeljeni na neuronskim mrežama mogu uspješno trenirati na dvjema razinama. U prvom se koraku kontinuirani vektori riječi uče upotrebljavajući jednostavne modele, a u drugom se koraku *n*-grami jezičnih modela uče na temelju distribuiranih reprezentacija riječi (Mikolov i sur., 2013a).



Slika 5: Model CBOW i model skip-gram su jednostavne neuronske mreže (A) koje se sastoje od ulaznoga, skrivenoga i izlaznoga sloja. U ulaznom i izlaznom sloju svaki čvor odgovara jednoj riječi. Tako je broj čvorova u slojevima jednak ukupnom broju leksikonskih jedinica modela. Broj čvorova skrivena sloja je jedan od parametara modela. Model trenira odnosno uči tako da pokretni prozor određene veličine prolazi kroz korpus i prilagođava težinski faktor tako da bi bolje odgovarao primjerima s kojima se susreće tijekom treniranja odnosno učenja. Ako prozor otkrije frazu poput "velik crn pas", model CBOW (B) reprezentira središnju riječ crn pobuđujući odgovarajući čvor u izlaznom sloju i sve su kontekstne riječi (velik i pas) istodobno pobuđene u ulaznom sloju. U modelu se skip-gram (C) asocijacijama između svake kontekstne riječi (velik i pas) predviđa ciljna riječ crn u zasebnom koraku učenja. Prilagođeno prema Manderi, Keuleersu i Brysbaertu (2017).

Prvi model koji su predložili Mikolov i sur. (2013a) nazvali su kontinuirani multiskup riječi (*Continuous Bag-of-Words Model*, CBOW). Sličan je jezičnim modelima temeljenima na unaprijednim neuronskim mrežama, samo što je uklonjen nelinearan skriven sloj, a projekcijski je sloj zajednički svim riječima. To znači da su vektori projicirani na isto mjesto. Mikolov i sur. (2013a) ovaj su model nazvali kontinuiranim multiskupom riječi jer poredak riječi nije bitan, a kontinuiran je zbog toga što se koristi kontinuiranim distribucijskim

reprezentacijama konteksta. CBOW, dakle, pokušava na temelju kontekstnih riječi odrediti središnju riječ.

Drugi model u okviru *word2vec*-a jest kontinuirani skip-gram model. Skip-gram je tehnika kojom se često koristi u obradi jezika. Pri oblikovanju n -grama moguće je "preskočiti" (*skip*) susjedne pojavnice (Guthrie i sur., 2006). Za razliku od modela CBOW ovaj model treba predvidjeti riječi koje prethode odnosno koje slijede središnju riječ na određenoj udaljenosti od nje. Budući da udaljenije riječi ne doprinose toliko značenju središnje riječi, tj. manje su semantički povezane s njom, težinski im je faktor manji (Mikolov i sur., 2013a). Kako bi učinkovitost modela skip-gram bila veća, Mikolov i sur. (2013b) predlažu uporabu hijerarhijskoga softmaxa kao aktivacijske funkcije. Njegova je prednost u tome što je to računalno učinkovitiji način procjene ukupne distribucije vjerojatnosti neuronskih čvorova u izlaznom sloju tako što se ne moraju valorizirati svi izlazni čvorovi leksikona W , već samo $\log_2(W)$ (Baroni, Dinu i Kruszewski, 2014; Mandera, Keuleers i Brysbaert, 2017; Mikolov i sur., 2013b). Alternativa je hijerarhijskom softmaxu negativno uzorkovanje (*negative sampling*). Mikolov i sur. (2013b) predlažu poduzorkovanje frekventnih riječi (*subsampling of frequent words*). To se posebno odnosi na uklanjanje zaustavnih riječi (*stop words*) kojih u tekstovima ima vrlo mnogo, a ne pridonose informativnosti. Zbog toga se u *word2vec*-u smanjuje njihov učinak tako što se vjerojatnost pojavljivanja svake riječi računa sljedećom formulom:

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$$

Pri tome je $P(w_i)$ vjerojatnost pojavljivanja riječi w_i , $f(w_i)$ je frekvencija riječi w_i , a t je prag. Najčešće je to prag od 10^{-5} . One riječi koje premašuju prag isključuju se. Pokazalo se da ova heuristički odabrana formula dobro funkcionira u praksi.

Baroni, Dinu i Kruszewski (2014) utvrdili su da *word2vec* nadmašuje sve dotadašnje modele, a posebice one koji se temelje na prebrojavanju. Goldberg i Levy (2014) su pokušali otkriti zašto *word2vec* može tako dobro reprezentirati riječi, no nisu uspjeli. Međutim, u svojim daljnjim istraživanjima (Levy i Goldberg, 2014; Levy, Goldberg i Dagan, 2015) utvrdili su da model skip-gram s negativnim uzorkovanjem, koji su zbog dobrih rezultata preporučili Mikolov i sur. (2013b), nije model koji se temelji na predviđanju, već na prebrojavanju. Otkrili su, naime, da je model skip-gram s negativnim uzorkovanjem zapravo implicitno rastavljanje na faktore pomaknute matrice točkaste procjene uzajamne informacije

(*shifted PMI matrix - pointwise mutual information*). Nadalje ističu da su neki parametri u modelu *word2vec* jasno obilježeni kao podesivi, dok to drugi nisu premda se i oni mogu podesiti. Tako Levy, Goldberg i Dagan (2015) razlikuju parametre koji se mogu podesiti prije, tijekom i nakon obrade. Ističu kako parametri koji se podešavaju prije same obrade utječu na ulazne podatke. Tako se, primjerice, poduzorkovanjem frekventnih riječi u istom prozoru oko središnje riječi nalaze riječi koje se ne bi ondje nalazile. Na taj se način povećava kontekstni prozor jer se uključuju i one riječi koje nisu bile obuhvaćene izvornim prozorom. Levy, Goldberg i Dagan (2015) takav oblik poduzorkovanja nazivaju "prljavim". Isto tako uklanjanje rijetkih riječi utječe na to koje će se riječi naći u kontekstom prozoru. No pokazalo se da taj parametar ne utječe mnogo na model.

Među parametre koji utječu na rezultate modela tijekom same obrade može se ubrojiti zaglađivanje kontekstne distribucije (*context distribution smoothing*). Ono smanjuje utjecaj točkaste procjene uzajamne informacije prema slabo frekventnim riječima.

Parametri koji se mogu podesiti nakon obrade jesu, primjerice, dodavanje kontekstnih vektora (*adding context vectors*), postavljanje težinskih faktora na svojstvene vrijednosti (*eigenvalue weighting*) i normalizacija vektora (*vector normalization*). Dodavanje kontekstnih vektora znači prema Levyu, Goldbergu i Daganu (2015) dodavanje pojmovna prvoga reda funkciji drugoga reda. Pojmovi prvoga reda mjere tendenciju da se jedna riječ pojavi u kontekstu neke druge riječi, a pojmovi drugoga reda mjere razinu do koje su te dvije riječi zamjenjive na temelju svoje tendencije da se pojavljuju u istom kontekstu. Potonje opisuje Harrisovu distribucijsku hipotezu (Harris, 1954).

Pomoću SVD-a deriviraju se vektori za riječi i kontekst ($W^{SVD}=U_d+\Sigma_d; C^{SVD}=V_d$). Istraživanja su pokazala da postavljanje težinskoga faktora na svojstvene vrijednosti matrice Σ_d s eksponentom p može imati značajan učinak na učinkovitost modela ($W^{SVDp}=U_d \cdot \Sigma_d^p$) (Bullinaria i Levy, 2012). Eksperimentirajući s četirima različitim vrijednostima za normalizaciju vektora (bez normalizacije, s normalizacijom redaka, s normalizacijom stupaca i kombinacija normalizacije redaka i stupaca), Levy, Goldberg i Dagan (2015) došli su do spoznaje da najbolje rezultate daje standardna normalizacija redaka.

Levy, Goldberg i Dagan (2015) u zaključku svojega rada uspoređuju model skip-gram s modelom CBOW. Ističu da se CBOW ne može svesti na rastavljanje na faktore matrice kojoj su redci riječi, a stupci konteksti. Model, naime, povezuje pojavnice koje se nalaze unutar kontekstnoga prozora predstavljajući kontekstne vektore kao zbroj vektora koji

reprezentira riječi. Na taj je način bolji u deriviranju reprezentacija riječi tako što kombinira pojavnice unutar kontekstnoga prozora.

Naposljetku Levy, Goldberg i Dagan (2015) zaključuju da modeli koji se temelje na predviđanju poput *word2vec*-a nisu bolji od modela koji se temelje na prebrojavanju. Jedina je razlika u podešavanju parametara o čemu se nije vodila dovoljna briga u modelima koji se temelje na prebrojavanju. Mandera, Keuleers i Brysbaert (2017) ne dijele takvo mišljenje ističući da su modeli koji se temelje na predviđanju mnogo kompaktniji u pogledu računalnih resursa bez obzira na to što moguće nisu bolji od modela koji se temelje na prebrojavanju. Isto tako ističu da referentne vrijednosti za računalnu lingvistiku nisu jednake onima u psiholingvistici. Nadalje ističu da osnova na kojoj počiva ideja modela koji se temelji na predviđanju jest ideja implicitnoga učenja. Pokazalo se kako je predviđanje riječi pomoću asociiranih riječi psihološki utemeljenije nego izrada matrice na kojoj se zatim primjenjuju razne transformacije i sl. Takav implicitan način učenja podudaran je s drugim modelima asocijativnoga učenja poput onoga koji su iznijeli Rescorla i Wagner (1972). Ovaj se model klasičnoga uvjetovanja temelji na asocijativnoj snazi između uvjetovanih i neuvjetovanih podražaja (*conditioned/unconditioned stimuli*). Drugim riječima, jaka asocijacija između uvjetovanih i neuvjetovanih podražaja znači da uvjetovan podražaj predviđa neuvjetovan. Model Rescorle i Wagnera pokušava objasniti kako se mijenja asocijativna snaga uvjetovanoga podražaja tijekom učenja s pojavom novih uvjetovanih podražaja sa svakim sljedećim izlaganjem podražaju (Miller, Barnet, Grahame, 1995). Baayen i sur. (2011) su ovaj model prenijeli na područje psiholingvistike. Prema njima asocijativna snaga znaka (*cue*) (u ovom su to slučaju slova) u odnosu na rezultat (*outcome*) (u ovome je to slučaju značenje u širem smislu: značenje riječi poput značenja riječi *računalo*, inflektivno značenje poput značenje pojedinih morfema koje sa sobom nose pojedini padeži i afiksarno značenje poput značenja različitih prefikasa ili sufikasa) jača je ako se usporedno pojavljuju. Razina do koje se asocijativna snaga može podesiti ovisi o broju drugih znakova. Ako je prisutno više znakova istovremeno, pozitivno je podešavanje manje, dok je negativno podešavanje veće i obrnuto. Drugim riječima, što se ispitanik češće susreće s jednom riječi, to će se s vremenom pojačati asocijativna snaga.

Model koji su predložili Rescorla i Wagner poseban je oblik povratnoga postupka (*backpropagation*) sa stohastičnim gradijentnim spustom (*stochastic gradient descent*).

Upravo na tim dvama mehanizmima za učenje počivaju modeli koji se temelje na predviđanju (Mandera, Keuleers i Brysbaert, 2017).

7. KORPUS

U proučavanju frekvencije riječi i općenito u drugim istraživanjima u psiholingvistici sve se više koristi korpusima. Korpusi su skup tekstova i predstavljaju bihevioralne podatke nastale jezičnim djelovanjem ljudi (Keuleers i Balota, 2015). Mogu poslužiti tome da bi se ustanovila učestalost pojavljivanja pojedinih riječi u jeziku. Tom se mjerom može objasniti najviše učinaka u leksičkom prepoznavanju. To znači da je frekvencija pojavljivanja pojedinih riječi odgovorna za brzo i učinkovito razlikovanje riječi od pseudoriječi. Tako, primjerice, Brysbaert i sur. (2011), pozivajući se na niz drugih istraživanja, tvrde da su ostali čimbenici koji se uobičajeno istražuju u psiholingvistici odgovorni za manje od pet posto varijance ako se izuzme frekvencija i ortografska bliskost.

Korpusi su temelj svake obrade u DSM-u jer su primarni izvor informacija o distribucijama riječi (Lenci, 2008). Zaključci do kojih se dolazi u cijelosti se temelje na korpusu. Algoritmi DSM-a koje smo opisali u prethodnom poglavlju propisane su procedure analize korpusa koje mogu biti različite. U DSM-u postoje dvije bitne odrednice koje mogu utjecati na rezultat koji se dobiva iz pojedinoga korpusa, a to su (i) model DSM-a koji se odabere i (ii) odabrani statistički model kojim se koristi za analizu dobivenih podataka.

Razilaze se, međutim, mišljenja oko toga kako konstruirati dobar korpus. Istraživanja su uglavnom provedena na engleskom govornom području i mjerodavan korpus je dugo vremena (pa čak i do danas) ostao korpus koji su konstruirali Francis i Kučera (1967). Unatoč tome što se pokazao nedostatnim jer, primjerice, ne odražava frekvencijske efekte, još je dugo vremena ostao u upotrebi. Na tu su činjenicu upozorili Brysbaert i New (2009). Uočili su da je korpus često konstruiran od raznih novinskih članaka, knjiga i časopisa. Jezik i konstrukcije koje se rabe u njima ne odražavaju onaj jezik kojim se služe ljudi u svakodnevnoj komunikaciji. To je jezik koji se temelji na raznolikosti i istančanom izrazu. Kako se u psiholingvističkim istraživanjima vrlo često želi odrediti vrijeme reakcije na pojedine ciljne riječi, frekvencija je dobar pokazatelj za takve učinke. Stoga su Brysbaert i New (2009) izradili nekoliko smjernica kojima bi se valjalo voditi prilikom konstrukcije korpusa za predviđanje frekvencije riječi.

1. Podslovi popularnih filmova i televizijskih serija dobar su izvor za konstrukciju korpusa jer se u njima rabi jezik koji je blizak ljudima i njihovoj svakodnevnoj međusobnoj komunikaciji. Polazi se od toga da su ispitanici češće gledali televiziju nego što su čitali knjige (New i sur., 2007; Brysbaert i sur., 2011).

2. Udžbenici za djecu iz osnovne i srednje škole važan su izvor za konstrukciju korpusa jer su česti ispitanici u psiholingvističkim istraživanjima upravo dodiplomski studenti. Oni su veći dio svojega dotadašnjega života proveli čitajući i učeći iz takvih udžbenika. Osim toga pokazalo se da riječi koje su usvojene ranije imaju stanovitu prednost u leksičkoj obradi nego one koje su usvojene kasnije (Izura i sur., 2011).
3. Tradicionalni pisani izvori poput knjiga također mogu doprinijeti poboljšanju korpusa. To je posebice slučaj ako se radi o popularnim štivima. Brysbaert, Keuleers i New (2011) u svojem su istraživanju, primjerice, pokazali da na ispitanike utječe vrijeme kada je neka knjiga izašla. Tako se pokazalo da stariji ispitanici brže reagiraju na riječi koje se pojavljuju u starijim naslovima nego mlađi ispitanici. Najbolje su se, međutim, pokazale one knjige koje su izdane relativno nedavno u odnosu na vrijeme kada je provedeno istraživanje.

Valja napomenuti da su u međuvremenu korpusi koji su konstruirani na temelju tekstova s društvenih mreža poput Facebooka i Twittera te različitih blogova pokazali bolje rezultate u odnosu na one koji su konstruirani na temelju serija i filmova (Heraĝdelen i Marelli, 2017; Gimenes i New, 2016). Jezik je društvenih mreža u odnosu na jezik koji je upotrijebljen u knjigama i podslovima spontaniji i nije korigiran ni lektoriran. Heraĝdelen i Marelli (2017) pretpostavljaju da je jezik društvenih mreža bolji u predviđanju frekvencijskih učinaka jer su tekstovi kraći. Veći broj kraćih jezičnih uzoraka te češće ponavljanje idiosinkratičnih izraza uzrok su toj činjenici. U svojem su istraživanju uočili da u korpusu SUBTLEX, sačinjenoga od podslova filmova i serija, prevladavaju izrazi koji se često upotrebljavaju u međusobnoj komunikaciji (lične zamjenice, oslovljavanje, odmilice i sl.), dok u korpusu koji je konstruiran od statusa s Facebooka i tweetova prevladavaju neslužbeni izrazi (*hubby, mom, auntie...*), izrazi osjećaja i riječi koje se odnose na neposrednu sadašnjost ili budućnost (rođendan, vikend, posao...). Ističu kako frekvencije riječi društvenih mreža mogu povećati varijancu za više od tri posto. Osim toga ukazuju na mogućnost da su frekvencije riječi na društvenim mrežama čvrsto povezane s prevalencijom riječi, tj. s time kojem je broju ljudi određena riječ poznata (Keuleers i sur., 2015). Internet se tako pokazao dobrim izvorom za konstrukciju korpusa. S jedne se strane pomoću njega može prilično brzo i jednostavno skupiti dovoljan broj tekstova, a s druge je strane jezik korišten na internetu raznovrsniji.

Mandera i sur. (2015) pokazali su kako dobro konstruiran korpus koji se temelji na uobičajenim tekstovima različitih izvora može imati jednako dobre rezultate kao i korpus koji je konstruiran od filmskih i serijskih podslova. Drugim riječima, jezični registar koji služi kao osnova sastavljanju korpusa bitno utječe na rezultate istraživanja.

Na rezultate istraživanja osim toga utječe i veličina korpusa. U velikom se broju članaka može pročitati da je bolji što veći korpus. Tako, primjerice, Recchia i Jones (2009) ističu da je važniji velik korpus od sofisticiranoga algoritma. Tome se protive De Deyne, Verheyen i Storms (2015). Oni tvrde da veličina korpusa nije ključna pri oblikovanju mentalnih reprezentacija. Na tom su tragu i Brysbaert i New (2009) kad zaključuju da je korpus od 16 do 30 milijuna pojavnica sasvim dostatan za pouzdane frekvencije. Ističu nadalje da se vrlo malo dodatne informacije može dobiti s korpusom većim od 30 milijuna pojavnica. Isto tako postoji mogućnost da se s većim korpusom težište premješta na teme koje nisu poznate velikom broju ljudi (Mandera, Keuleers i Brysbaert, 2017). Na taj se način mogu obuhvatiti i one riječi koje se pojavljuju manje od jednom na milijun riječi (Brybaert, Mandera i Keuleers, 2017). Drugim riječima, optimalna veličina korpusa ovisi o svrsi za koju se korpus želi upotrijebiti. Ako je, primjerice, cilj istraživanja odrediti vrstu riječi niskofrekventnim riječima ili dobivanje informacija o usporednom pojavljivanju riječi, potreban je veći korpus. Veliki korpusi tako polučuju dobre rezultate na testovima poznavanja jezika poput TOEFL-a i drugih, no za psiholingvistička istraživanja nisu toliko pogodna. Kako su zadaci semantičkoga usmjeravanja prilično česti u takvim istraživanjima, raznolikost riječi nije toliko bitna kao što je to u testovima poznavanja riječi. Upravo je ta jezična raznolikost bitna odrednica velikih korpusa. Isto tako valja napomenuti da korpusi koji su iznad određenoga broja pojavnica više nisu smisleni i vjerodostojni. Prototipski je ispitanik student na dodiplomskom studiju. Ako se polazi od toga da s 18 godina kreće na studij i da svaki dan od rođenja čita 16 sati po 300 pojavnica u minuti, za vrijeme svojega života mogao je pročitati najviše 1.89 milijardi pojavnica (Mandera, Keuleers i Brysbaert, 2017). To znači da svaki korpus veći od tog broja pojavnica nije suvisao i nadilazi ljudske mogućnosti. Međutim, taj je broj također preuveličan jer je nemoguće da čovjek pročita toliko mnogo, pogotovo ne od rođenja.

U posljednje se vrijeme pojavljuju korpusi koji mogu biti iznimno veliki. Takvi su skupovi podatak (*dataset*) veliki i do bilijun riječi (10^{12}). Tako se veliki korpusi mogu obraditi

samo s posebnim modelima poput modela GPT-3 (Brown i sur., 2020). Takav je model, međutim, vrlo velik i stoga energetski vrlo zahtjevan. Osim toga je nedostupan.

Zaključno se može reći da dobro izbalansiran korpus nastao na temelju internetskih stranica može jednako dobro poslužiti u psiholingvističkim istraživanjima kao i oni korpusi nastali na temelju podslova, blogova, tweetova i zapisa s Facebooka. Valja pri tome voditi računa da je korišten korpus primjerene i suvisle veličine te da je adekvatnoga registra. Ako DSM-u stoji na raspolaganju takav korpus, mogu se ostvariti željeni rezultati. Ovisno o korištenom DSM-u veličina prozora može predstavljati ograničenja čovjekove radne memorije, a veličina korpusa i njegova kvaliteta određuju neposredovano ljudsko iskustvo (Brysaert i sur., 2011). Brysaert, Mandera i Keuleers (2017) upozoravaju da u radu s velikim korpusima valja biti vrlo oprezan jer mnoge stvari mogu krenuti po zlu zbog čega daju nekoliko savjeta: (i) treba redovito provjeravati izračune, (ii) treba upotrijebiti dokazane algoritme jer nisu svi algoritmi jednako dobri i (iii) valja voditi brigu o kvaliteti korištenoga korpusa (to se često opisuje frazom "*garbage in, garbage out*").

8. MODELI DVOJEZIČNOGA MENTALNOG LEKSIKONA

Kako Erdeljac (2009) tvrdi, mentalni je leksikon "skladište" riječi sa svim ključnim fonetskim, fonološkim, morfološkim, sintaktičkim i semantičkim obilježjima. Pojedinci se aktivno koriste njime svaki dan tako da su podaci u njemu uvijek aktualni. Neprestano se širi jer pojedinac uči i usvaja nove riječi. Mnogi su istraživači pokušali opisati kako je konstruiran mentalni leksikon. Teorije koje su se razvile na području jednojezičnoga istraživanja poslužile su kao temelj onima na dvojezičnom.

Prvo je pitanje kojim su se istraživači bavili proučavajući dvojezični mentalni leksikon jesu li leksikoni dvaju jezika koje pojedinac govori međusobno odvojeni ili su zajednički. Drugim riječima, jesu li jedinice leksikona pohranjene u dva odvojena skladišta ili pak u jednom zajedničkom. Različitim su eksperimentima u brojnim istraživanjima pokušali shvatiti i opisati organizaciju mentalnoga leksikona. Danas među istraživačima prevladava mišljenje da je mentalni leksikon dvaju jezika dvojezičnoga govornika zajednički (za pregled vidi de Groot, 2011).

Drugi problem na koji su se znanstvenici fokusirali jest pristup mentalnom leksikonu. Je li taj pristup selektivan ili neselektivan, odnosno jesu li uvijek istodobno aktivna oba leksikona pojedinca ili ne. Primjerice, Dijkstra je u suradnji s mnogim kolegama došao u nizu eksperimenata usmjeravanjem do zaključka da je pristup mentalnom leksikonu neselektivan pri čemu važnu ulogu igraju postavke zadatka koje ispitanici moraju ispuniti (Dijkstra, 2003; 2005; Dijkstra i sur., 2007; Dijkstra i van Heuven, 2002; van Heuven, Dijkstra i Grainger, 1998). Do toga je zaključka došao tako što je uočio da se kognati (riječi koje dijele istu ili sličnu ortografiju i značenje u dva ili više jezika) prepoznaju mnogo brže u zadatku leksičke odluke od onih riječi koji to nisu. Nadalje Kroll i Tokowicz (2005) potvrđuju činjenicu neselektivnosti istraživanjima na području mozga oslanjajući se pri tome na Illes i sur. (1999). Zaključile su, naime, da su semantički primitivi dostupni obama jezicima. Drugim riječima, riječi čije je značenje razumljivo samo po sebi i kojima se koristi pri objašnjavanju drugih riječi (Wierzbicka, 1996) postoje u oba leksikona dvojezičnoga govornika koji im pristupa na jednak način. Svoju tezu oslanjaju na spoznaji prema kojoj su aktivirana ista kortikalna područja u eksperimentima određivanja konkretnosti ispitanih riječi.

Na temelju svojih istraživanja i prethodnih modela jednojezične organizacije mentalnoga leksikona, istraživači dvojezičnoga leksikona izradili su svoje modele. Gulan

(2016) je u svojem radu dvojezične modele mentalnoga leksikona podijelila na hijerarhijske i konektivističke modele. Hijerarhijski se modeli odlikuju time što postoji stanovit redosljed pristupanja podacima, dok je za konektivističke modele karakteristično da se oslanjanju na način rada neuronskih sustava. Tako među hijerarhijske modele ubraja modele konceptualne i asocijativne organizacije (Potter i sur., 1984), revidirani hijerarhijski model (*Revised Hierarchical Model* - RHM) (Kroll i Stewart, 1994; Kroll i sur., 2010) i inhibicijski kontrolni model (*Inhibition Control* - IC) (Green, 1993; Green, 2000; Green, 2010). U konektivističke modele ubraja bilingvalni interaktivni aktivacijski model (*Bilingual Interactive Activation* - BIA) (Dijkstra i van Heuven, 1998) i njegova kasnije proširena verzija BIA+ (Dijkstra i van Heuven, 2002), dvojezični interaktivni model leksičkoga pristupa i slušnoga prepoznavanja riječi (*Bilingual Interactive Model of Lexical Access* - BIMOLA) (Léwy i Grosjean, 2008) i model distribucijskih konceptualnih obilježja (*Distributed Conceptual Feature Model* - DCFM) (de Groot, 1992a; de Groot, 1992b; de Groot, 1993).

Odlika je konektivističkih modela ta da se konektivistička mreža sastoji od nekoliko slojeva. Jedinice mreže međusobno se aktiviraju kada prijeđu određen prag, tzv. razinu aktivacije. Mreža se sastoji od tri sloja: sloj unosa (*input layer*), sloj izlaza (*output layer*) i skrivenoga sloja koji povezuje prethodna dva sloja (*hidden layer*) (Cergol, 2011). Thomas i van Heuven (2005) u svojem radu dijele konektivističke modele dalje na lokalističke i distribucijske modele. Dok lokalistički modeli objašnjavaju kako funkcioniraju strukture mentalnoga leksikona zadužene za obradu podataka kod odraslih dvojezičnih govornika, distribucijski se modeli uglavnom bave problemima usvajanja jezika. Lokalistički modeli, kako Thomas i van Heuven (2005) ističu, pridijevaju diskretne entitete individualnim jedinicama. Pojednostavljeno, aktivacija se širi uzduž veza između čvorova usmjerivača i ciljne riječi u konektivističkoj mreži. Postavke se tih modela određuju na početku tako da se ne mijenjaju tijekom vremena. Drugačije rečeno, odlike modela postavlja inicijalno istraživač i one se ne mijenjaju tijekom učenja, tj. tijekom procesa u kojem model stječe nova iskustva. Za razliku od takvih modela, distribucijski modeli predstavljaju pojedinačne entitete kao ustaljene načine širenja aktivnosti u nizu jedinica konektivističke mreže. U semantičkom usmjeravanju to znači da reprezentacija usmjerivača u mentalnom leksikonu izravno pobuđuje temeljne semantičke čvorove koje dijeli s reprezentacijama ciljne riječi u mentalnom leksikonu. Takvi se modeli mijenjanju kako model uči i stječe nova iskustva. Početne su postavke u tim modelima nasumično postavljene tako da se mogu mijenjati tijekom procesa

učenja. U suprotnom model ne bi ništa naučio i ostao bi nepromijenjen. Prema tome, model počinje i bez kakvog sadržaja. Tek u procesu učenja koji slijedi povezuje se, primjerice, oblik riječi sa značenjem. Thomas i van Heuven tako u lokalističke modele ubrajaju modele BIA i BIA+ te međustupanj između ta dva modela, a to je model semantičke, ortografske i fonološke interaktivne aktivacije (*Semantic, Orthographic, and Phonological Interactive Activation* - SOPHIA), dok u distribucijske modele ubrajaju dvojezični model jedinstvene mreže (*Bilingual Single Network Model* - BSN) (Thomas, 1997), model dvojezične jednostavne rekurentne neuronske mreže (*Bilingual Simple Recurrent Network* - BSRN) (French, 1998) i samoorganizacijski model dvojezične obrade (*Self-Organizing Model of Bilingual Processing* - SOMBIP) (Li i Farkaš, 2002).

Budući da će se u ovome radu prije svega obrađivati razumijevanje pisane riječi kod dvojezičnih ispitanika, a ne kako će se odvijati njihov jezični razvoj, u obzir ćemo uzeti poglavito lokalističke modele mentalnoga leksikona.

U nastavku ćemo predstaviti neke od najvažniji lokalističkih konektivističkih modela dvojezičnoga mentalnog leksikona. Osvrnut ćemo se i na preteče modela koji su danas najčešće u uporabi te ocrtati njihove glavne odlike.

8.2. Inhibicijski kontrolni model (IC)

Ovaj je model u nekoliko članaka razradio David W. Green (2000a pretisak iz 1986; 1998; 2000b; 2011; von Studnitz i Green, 2002). Iako je u prvom redu usmjeren na produkciju, može se primijeniti i na razumijevanje jezika. U svojem temeljnom radu Green (2000a, pretisak iz 1986) ovaj model zasniva na tri ključne ideje do kojih je došao u radu s afazičnim pacijentima. To su (i) kontrola, (ii) aktivacija i (iii) resursi. Polazi od činjenice da su afazični dvojezični govornici ponekada u stanju razumjeti jezik, ali ga ne mogu govoriti. Green zaključuje da su podsustavi za razumijevanje i proizvodnju jezika međusobno odvojeni. Green nadalje zaključuje da i nakon prestanka rada jednoga podsustava drugi može nastaviti raditi.

Kako bi se određena riječ iz mentalnoga leksikona mogla aktivirati, potrebno je prijeći prag aktivacije. Green polazi od toga da se proces proizvodnje riječi može dalje podijeliti u dva koraka. U prvome pojedinač aktivira značenje riječi, dok u drugom aktivira zvučnu sliku odnosno fonološki oblik riječi. Prema tome, Green pretpostavlja postojanje stanovitoga unutarnjeg preklopnika (*switch*) koji može u cijelosti aktivirati odnosno deaktivirati jedan od

dvaju jezika dvojezičnoga govornika. Međutim, pokazalo se da je jezik koji nije odabran i dalje aktivan, odnosno nije deaktiviran do te mjere da pojedinac mentalnom leksikonu neodabranoga jezika ne može više pristupiti. Ukoliko se pak neodabrani jezik često ne koristi ili ga se zanemari, može doći do toga da njegova razina aktivacije raste. Drugim riječima, teže se aktiviraju riječi iz toga jezika. Green razlikuje tri razine aktivacije jezičnoga sustava: (i) odabran (jezik koji kontrolira produkciju), (ii) aktivan (jezik koji igra određenu ulogu u trenutačnoj produkciji) te (iii) prikriven (jezik koji se nalazi u dugoročnom pamćenju, ali ne igra nikakvu ulogu u trenutačnoj produkciji). Razlog je aktivacije određenoga suspregnutog jezika taj što se dvojezični govornici često njime koriste u svakodnevnom životu.

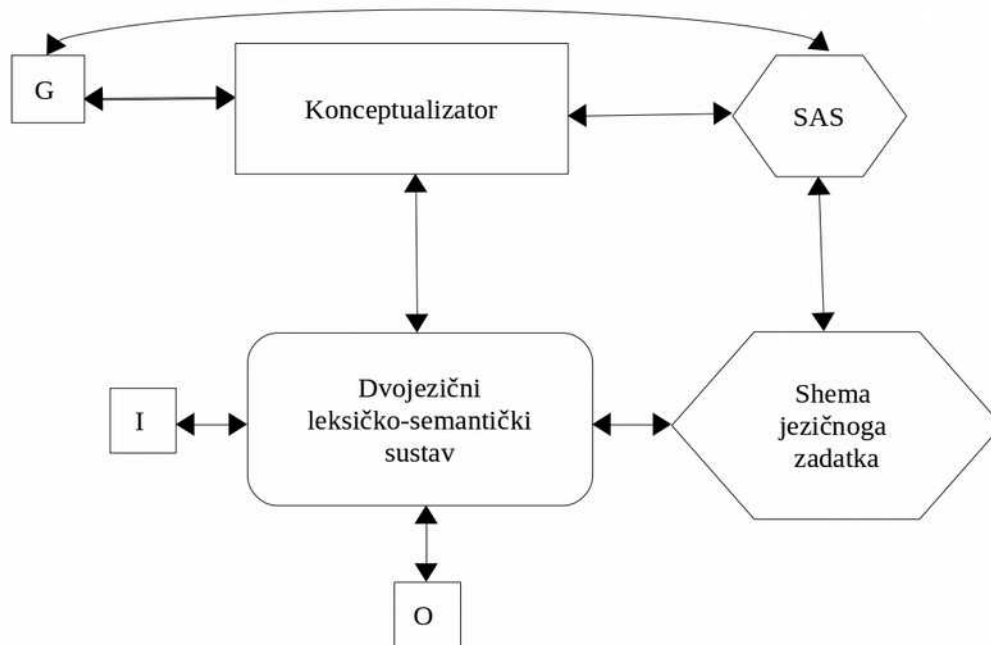
Green smatra da moraju postojati nekakvi mehanizmi koji aktiviraju odnosno potiskuju aktivnost pojedinoga jezika. To se postiže resursima koji mogu pobudi sustav ili ga suspregnuti (*excitatory resource* i *inhibitory resource*). Svaki oblik kontrole iskorištava određene resurse, tj. energiju kako bi mogao raditi.

Kako bi dvojezični govornik uopće mogao govoriti na jednom od svoja dva jezika, Green uvodi uređaj koji naziva specifikator (*specifier*). Zadužen je za kontrolu jezičnoga sustava kako bi pojedinac govorio samo na željenom jeziku. To znači da sustav mora biti tako podešen da su aktivni oni podsustavi za prepoznavanje i proizvodnju onoga jezika koji je odabran, a potisnuti oni podsustavi jezika koji to nije. Podsustavi mogu biti aktivirani ili potisnuti iznutra i izvana. To znači, primjerice, da je J2 suspregnut izvana ako pojedinac spontano govori J1, a ako prevodi s J2 na J1, J2 je suspregnut iznutra.

Važnu ulogu u Greenovu modelu igra shema jezičnoga zadatka (*language task schema*) (Green, 1998). Ona predstavlja niz mentalnih radnji koje se trebaju obaviti kako bi se izveo određen zadatak. Kod zadataka koje je govornik već prije obavljao, shema jezičnoga zadatka je automatizirana i pohranjena je u memoriji kojom se pojedinac može poslužiti u izvedbi zadataka. Međutim, ako je suočen sa zadatkom koji do tada još nije obavljao, u pomoć dolazi sustav nadziranja pažnje (*supervisory attentional system* - SAS). On nadzire rutinsko obavljanje shema zadataka i uključuje se po potrebi. Sustav SAS uključuje točno određenu shemu jezičnoga zadatka koja potom kontrolira ponašanje tako što mijenja razinu aktivacije lema.

Za IC model je nadalje bitan sustav koji je Green nazvao konceptualizatorom (*conceptualizer*). To je izvanjezični sustav koji konstruira konceptualne reprezentacije iz informacija pohranjene u dugoročnom pamćenju. Ključan pojam u IC modelu jest jezična

lema koja sadrži jezične oznake o tome kojem jeziku pripada pojedina riječ. IC model zorno je prikazan na Slici 6.



Slika 6: Inhibicijski kontrolni model (IC) (Green, 1998). Konceptualizator oblikuje konceptualne reprezentacije koje se temelje na informacijama u dugoročnom pamćenju te potiče određene ciljeve posredstvom jezika. Proces njihova postizanja nadgleda sustav nadziranja pažnje (supervisory attentional system – SAS), leksičko-semantički sustav i shema jezičnoga zadatka. Shema jezičnoga zadatka kontrolira izlaz iz leksičko-semantičkoga sustava. Da bi se izrekla pojedina riječ, SAS prenosi specifikacije određenoga jezika shemi jezičnoga zadatka, a konceptualizator prenosi konceptualne informacije jezično-semantičkom sustavu. Prilagođeno prema Greenu (1998).

De Groot (2011) razlikuje dvije bitne odrednice po kojima se IC model razlikuje od drugih modela jezične kontrole. To su (i) sustav SAS koji određuje kako treba podesiti jezični sustav da se izvrši određen zadatak te (ii) potiskivanje kontekstno neprimjerenih leksičkih jedinica. To znači da će lema biti to više potisnuta što je ona aktivnija na neciljnom jeziku. Tako će, primjerice, u prevođenju potiskivanje biti jače ako je J2 izlazni jezik nego kada je to J1. J1 se, naime, mora više potiskivati jer je J2 često slabiji u odnosu na J1. De Groot tako zaključuje da se kontrola može vršiti proaktivno odnosno globalno tako da se unaprijed prilagodi razina aktivacije lema u oba jezika te reaktivno odnosno lokalno potiskujući aktivaciju lema neodabranoga jezika da bi prošle pokraj kontrolnoga mehanizma.

Green (2011) ističe da je u svojim istraživanjima došao do spoznaje da je jezična kontrola zapravo samo jedan oblik općega sustava kontrole. To znači da je sustav jezične kontrole samo podsustav općega sustava kognitivne kontrole koji regulira i nadzire uz jezične i druge kognitivne procese.

Rad von Studnitz i Greena (2002) snažno je utjecao na BIA+. Posebno su se Dijkstra i van Heuven (2002) vodili idejom tzv. izvanjske kontrole. To znači da uvjeti koje pred ispitanika postavlja zadatak mogu utjecati na ishod eksperimenta. Zbog toga su Dijkstra i van Heuven u svoju proširenu verziju modela BIA+ uključili i shemu zadatka.

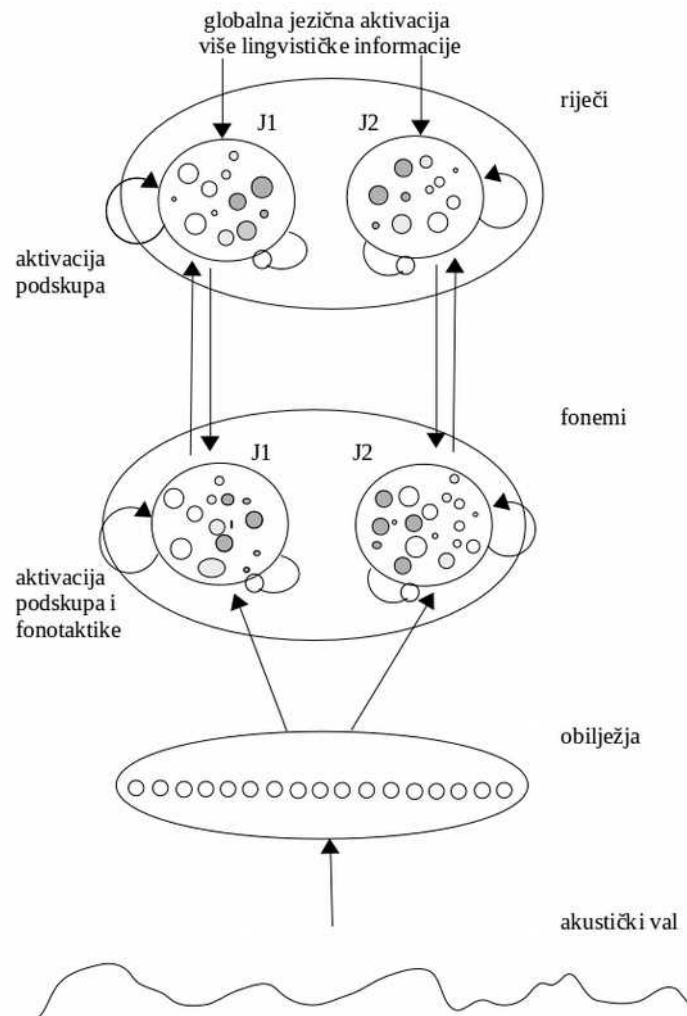
8.2. Dvojezični interaktivni model leksičkoga pristupa i slušnoga prepoznavanja riječi (BIMOLA)

Léwy i Grosjean (2008) razvili su ovaj računalni lokalistički model kako bi se mogla proučiti percepcija govora. Puni naziv toga modela je *Bilingual Interactive Model of Lexical Access*. Ovaj se model temelji na načelima interaktivnoga aktivacijskog modela slušnoga prepoznavanja riječi TRACE (McClelland i Elman, 1986). Isto kao i TRACE BIMOLA ima slojeve za obilježja, foneme i riječi. Razinu obilježja dijele oba jezika, dok su slojevi za foneme i riječi specifični za pojedine jezike, odnosno organizirani su u odvojene jezične podsustave koji su pak dio većega sustava. Model BIMOLA polazi od toga da je aktiviran samo jedan jezik ako dvojezični govornik komunicira s jednojezičnim govornikom, dok su oba jezika aktivirana ako komunicira s dvojezičnim govornikom s kojim dijeli svoja dva jezika. Pri tome dvojezični govornik može miješati riječi iz obaju jezika. U tom slučaju jedan se jezik smatra osnovnim ili temeljnim jezikom, a drugi jezik jezikom gostom.

Ovaj model služi kako bi se objasnili sljedeći dvojezični učinci:

1. učinak temeljnoga jezika (riječ jezika gosta koja je u kontekstu jezika gosta nedvojbeno pripada jeziku gostu)
2. učinak sličnosti jedinica (pojedine se jedinice jezika poput fonema koje međusobno dijele oba jezika aktiviraju istodobno)
3. fonotaktički učinak (brže se prepoznaju riječi jezika gosta koje svojim fonotaktičkim obilježjima jednoznačno odaju pripadnost jeziku gostu)
4. fonetski učinak (postojat će veći broj kandidata u jeziku gostu u slučaju prebacivanja kodova (*code switching*) nego u slučaju da se samo posuđuju pojedine riječi iz drugoga jezika)

5. homofonski učinak temeljnoga jezika (otežano će se procesirati riječi jezika gosta u kojem postoje homofoni koji su vrlo slični riječima temeljnoga jezika)
6. međujezični učinak prepoznavanja temeljen na frekvenciji homofona (koliko će se jednostavno prepoznati homofoni jezika gosta ovisi o odnosu njegove frekvencije u jeziku gostu i temeljnom jeziku, odnosno otežano će se prepoznati one riječi jezika gosta koje imaju homofon u temeljnom jeziku)



Slika 7: Bilingvalni interakcijski model leksičkoga pristupa (BIMOLA) (Grosjean, 2008). Model se sastoji od triju razina: obilježja, fonema i riječi. Prvu razinu obilježja dijele oba jezika, dok su ostale dvije razine (fonema i riječi) istodobno i neovisne i povezane. Svaki je jezik predstavljen podskupom koji se sastoji od jedinica samo toga jezika, ali istodobno su oba jezika podskupovi većega skupa. Fonemi i riječi mogu imati bliže i dalje susjede što je prikazano prostornim odnosima i svjetlinom. Aktivacija može ići u smjeru odozdo prema gore (npr. fonemi pobuđuju riječi koji ih sadrže), odozgo prema dolje (temelje se na izvanjezičnim informacijama o slušateljevoj jezičnoj aktivaciji [language model]) i horizontalno (fonotaktička aktivacija koja se temelji na distribuciji fonema u inicijalnom dijelu riječi ili aktivacija podskupa koja se temelji na pojačavanju aktivacije drugih riječi istoga jezika ukoliko se određena riječ prepoznaje da pripada tom jeziku). U BIMOLA-i postoji samo lateralna inhibicija elemenata istoga podskupa unutar jednoga jezika. Aktivacijske su veze označene strelicama s trokutastom, a potiskivajuće veze onima s okruglom glavom. Prilagođeno prema Grosjeanu (2008).

Riječi određenoga jezika mogu se aktivirati unaprijed odozgo prema dolje što je uvjetovano izvanjezičnim utjecajima poput jezične aktivacije (*language mode*) ispitanika. Sustav je postavljen tako da su aktivacijske razine dvaju jezika unaprijed zadane odozgo prema dolje te se tek s vremenom mijenjaju pod utjecajem unutarjezičnoga pobuđivanja (vidi Sliku 7). Koji će pak jezik biti inicijalno pobuđeniji, ovisi o jezičnoj aktivaciji. Postoje, naime, dvije krajnje točke jezične aktivacije: To su jednojezična aktivacija, u kojoj dvojezični govornik isključivo govori samo jedan jezik, te dvojezična aktivacija, u kojoj govornik često mijenja jezike i prelazi iz jednoga u drugi. Između tih dviju točaka postoji čitav niz međustupnjeva. Valja, međutim, napomenuti da je kod dvojezične aktivacije temeljni jezik uvijek pobuđeniji nego jezik gost (Léwy i Grosjean, 2008). Kako postoje unaprijed zadani mehanizmi za pobuđivanje odozgo prema dolje, tako postoje i mehanizmi u kojima je smjer pobuđivanja odozdo prema gore. To znači da se određen osjetilni podražaj prenosi na sljedeću višu razinu. Na taj se način prepoznaje jezik kojim se trenutačno govori. Isto tako što se više riječi nekoga jezika prepoznaje, to će dotični podsustav biti aktivniji.

Prema de Groot (2011) model BIMOLA do sada je uspješno simulirao dva gore navedena učinka: učinak sličnosti jedinica i homofonski učinak temeljnoga jezika. Ističe nadalje da je model SOPHIA bolji utoliko što uključuje i vizualno prepoznavanje riječi. BIMOLA i BIA odnosno SOPHIA dijele mnoge sličnosti (za detalje o modelu SOPHIA vidi poglavlje 8.5.). Jedna je od njih ideja o međusobnom pobuđivanju čvorova, tj. čvorovi na određenoj razini u sustavu mogu pobuditi čvorove susjednih razina. Osim toga u svim navedenim modelima dolazi do lateralne inhibicije, samo što se ona kod modela BIMOLA ograničava na razinu obilježja. Za razliku od modela BIA u kojem može doći do lateralne inhibicije na razini fonema i riječi između dvaju jezika govornika, u modelu BIMOLA to je moguće samo unutar jednoga jezika. Drugim riječima, BIMOLA polazi od odvojenih mentalnih leksikona (Thomas i van Heuven, 2005).

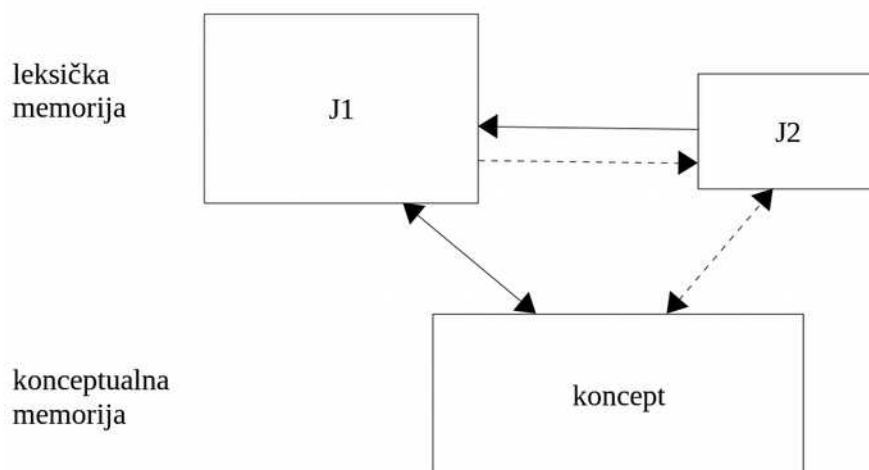
8.3. Revidirani hijerarhijski model (RHM)

Revidirani hijerarhijski model nastao je na temelju rada Kroll i Stewart (1994). Razvio se iz dvaju prethodnih modela: modela asocijacija riječi i modela konceptualne medijacije. Ta su dva modela predložili Potter i sur. (1984). Prvi polazi od toga da su riječi na J2 povezane asocijativnim vezama s riječima na J1 i samo se posredstvom J1 mogu povezati s konceptualnom memorijom. Konceptualna memorija je onaj dio mentalnoga leksikona u

kojem su pohranjeni podaci o značenju riječi, ali ne i podaci o njezinu obliku. Za razliku od toga u modelu konceptualne medijacije riječi na J2 imaju izravan pristup konceptualnoj memoriji.

Ti su se modeli pokušali dokazati eksperimentima prevođenja i imenovanja slika. Polazi se od toga da prema modelu asocijacija riječi prevođenje unaprijed (iz J1 na J2) ide brže od imenovanja slika na J2. Razlog tomu jest činjenica što za prevođenje nije potrebno posredovanje konceptata što je pak potrebno u slučaju imenovanja slika. Za razliku od toga model konceptualne medijacije predviđa da će prevođenje unaprijed biti jednako brzo kao i imenovanje slika na J2 zbog toga što se za obje radnje mora poslužiti konceptualnom memorijom prije nego što se može pristupiti riječima na J2.

Kroll i Stewart (1994) smatraju da se pojedinac u početnim stadijima učenja J2 u najvećoj mjeri oslanja na prijevode na svojem J1, tj. pristupa značenjima riječi na J2 posredno preko leksičkih veza. Kako njegova vrsnost na J2 raste, sve se više oslanja na konceptualne veze, premda leksičke nikad nisu u potpunosti deaktivirane. One nadalje ističu da su leksičke veze među leksičkim jedinicama pojedinoga leksikona te konceptualne veze između pojedinoga leksikona i konceptualne memorije dvosmjerne, no razlikuju se u jačini. Pretpostavlja se da je leksička veza od J2 do J1 jača od leksičke veze u obrnutom smjeru, od J1 do J2, jer se riječi na J2 prvo asociiraju s riječima na J1. Isto se tako pretpostavlja da je konceptualna veza između J1 i konceptualne memorije jača od one između J2 i mentalnoga leksikona (vidi Sliku 8). Kroll i Stewart zaključuju da je prevođenje unaprijed, iz J1 na J2, osjetljivo na semantičke i konceptualne informacije, dok to nije slučaj u prevođenju unazad, iz J2 na J1. Drugim riječima, put prevođenja unazad je kraći od puta prevođenja unaprijed jer nema dodatnoga koraka preko konceptualne memorije. De Groot (2011) smatra da u procesu prevođenju unaprijed treba više vremena zbog toga što je proces proizvodnje riječi na J2 mnogo zahtjevniji nego njezino prepoznavanje.



Slika 8: Revidirani hijerarhijski model (RHM) (Kroll i Stewart, 1994). Između J1 i J2 postoje izravne leksičke veze između prijevodnih ekvivalenata i obrnuto. Razlika je u tome što su one u smjeru J2-J1 jače nego u smjeru J1-J2. Jača je veza označena neprekinutom linijom strelice, dok je slaba označena isprekidanom linijom strelice. Postoje i razlike u jačini veza između pojedinoga leksikona i konceptualne memorije. Tako su veze između konceptualne memorije i J1 jače od onih s J2. Posljedica toga je razlika u vrsnosti između J1 i J2. Prilagođeno prema Kroll i Stewart (1994).

U svojim su eksperimentima Kroll i Stewart (1994) došle do spoznaje da do kategorijalne interferencije dolazi kod vrsnih dvojezičnih govornika samo ako prevode unaprijed. Kako je mentalni leksikon J2 manji, potrebno je više vremena da mu se pristupi.

Brysbaert i Duyck (2010) iznijeli su kritiku ovoga modela. Istaknuli su da je relativno malo dokaza o postojanju odvojenih leksikona za svaki jezik kojim se dvojezični govornik služi te da im se pristupa selektivno. Isto su tako utvrdili da uključivanje pobuđivačkih veza između prijevodnih ekvivalenata na leksičkoj razini može dovesti do toga da spriječi prepoznavanje riječi. Veze su između riječi na J2 i njihovih prijevoda prema njima jače nego što to predviđa RHM. Naposljetku smatraju da treba razlikovati semantička obilježja koja su ovisna o jeziku i ona semantička obilježja koja su neovisna o jeziku. Do posljednje su tvrdnje došli na temelju rada Paradisa (1997) koji je ustanovio da se na engleskom jeziku ne može reći *a large sister* kad se misli na stariju sestru, već se mora koristiti izrazom *a big sister*. S druge se strane bez pogreške može reći *a large book* i *a big book* ako se misli da debelu knjigu. Ta se činjenica pak ne može temeljiti na semantičkom sustavu koji je neovisan o jeziku. Isto su tako Sahlin, Harding i Seamon (2005) potvrdili da se u zadatku pamćenja ispitanici mnogo bolje prisjećaju riječi ako su na istom jeziku na kojem su pamtili riječi nego

kada su na drugom jeziku. Ta činjenica upućuje na to da se riječi nisu pamtile neovisno o jeziku.

U reakciji na kritiku ovoga modela, koju su iznijeli Brysbaert i Duyck (2010), Kroll i sur. (2010) potvrdili su da se vrlo rano pokazalo kako posredovanje J1 za razumijevanje riječi nije nužno potrebno. Isto su tako relativizirali tvrdnju o postojanju slabije veze između riječi na J2 i konceptualne memorije. Ustanovile su da slabija veza nije dvosmjerna jer se u nizu istraživanja pokazalo da je asimetrija između njih bitnija za proizvodnju riječi nego za njihovo prepoznavanje.

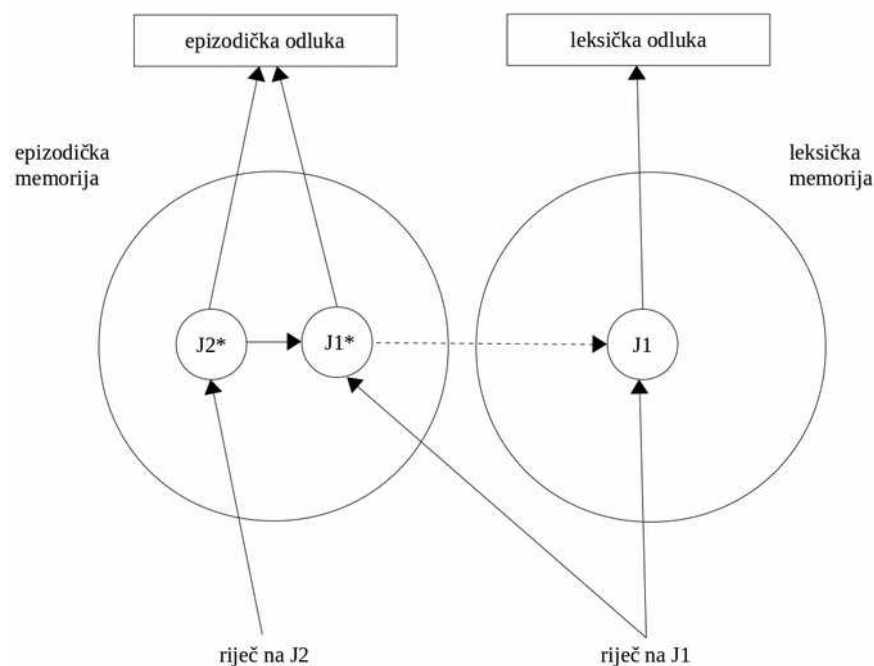
Osim u objašnjavanju produkcije riječi RHM-om se koristi i pri objašnjavanju recepcije riječi. Kroll i Sholl (1992) su tako primijetili da je učinak usmjeravanja u zadatku leksičke odluke s usmjeravanjem veći ako su usmjerivači na J1, a ciljne riječi na J2. Dodatan je uvjet i relativno kratko vrijeme do pojavljivanja ciljne riječi (*stimulus onset asynchrony* - SOA). Ako je SOA pak duži, pokazalo se da nema razlike u učinku usmjeravanja bilo da je usmjerivač na J1 ili na J2 (vidi de Groot, 2011). RHM postulira da ne bi trebalo doći ni do kakvoga učinka usmjeravanja ako je usmjerivač na J2, a ciljna riječ na J1 jer je konceptualno posredovana. Taj se učinak ipak pojavljuje. Ako je poredak obrnut, usmjerivač je na J1, a ciljna riječ na J2, učinak je veći. Takvo se pojavljivanje učinaka, osobito ako je SOA dulji, može protumačiti tako što su riječi na J2 imale više vremena da pristupe konceptualnoj memoriji kako bi pronašle potrebne informacije. Nasuprot tome, pri objašnjenju ovakve pojave može se uzeti u obzir tumačenje de Groot (1992a; 1992b) u okviru distribucijske reprezentacije značenja prema kojem konkretne riječi dijele niz konceptualnih čvorova u oba jezika. Drugim riječima, čim se prepozna usmjerivač, on pristupa svojim distribuiranim reprezentacijama značenja i na taj način uzrokuje ubrzano prepoznavanje ciljne riječi (de Groot, 2011). Može se, dakle, zaključiti da se ispitanici bez obzira na svoju vrsnost u poznavanju J2, oslanjaju na konceptualnu memoriju neovisno o tome prevode li unaprijed ili unazad.

Ovaj je model utoliko važan što jasno razlikuje leksičke i konceptualne reprezentacije u konceptualnoj memoriji te što prikazuje kako se jačina veza između njih mijenja s povećanjem vrsnosti u jeziku (Gulan, 2016).

8.4. Hipoteza o epizodičkom pamćenju

Ova hipoteza pretpostavlja da se pamćenje jasno dijeli na epizodičko i leksičko pamćenje. Prvo pohranjuje podatke o događajima, dok potonje pohranjuje jezične podatke o riječima (Jiang i Forster, 2001). Jiang i Forster ističu da se epizodičko pamćenje ne može shvatiti onako kako je to definirao Tulving (1972), tj. kao autobiografski podaci, već prije kao ne-leksičko pamćenje kako je to opisao Forster (1985) (Jiang i Forster, 2001; Witzel i Forster, 2012).

J1* predstavlja epizodički zapis riječi J1. J1* je, dakle, riječ smještena u kontekstu, dok je J1 njezin apstraktni zapis koji nije vezan za određen kontekst. Veza je između J1* i J1 višestruko potvrđena učestalim pojavljivanjem epizodičkoga zapisa riječi J1* u različitim kontekstima. Riječi na J2, s druge strane, nemaju leksički zapis, već postoje samo u epizodičkom pamćenju (J2*). Takvi su epizodički zapisi J2* asocijativno povezani s istim takvim epizodičkim zapisima na J1*. Tek epizodički zapis na J1* može aktivirati značenje na J1. Prikaz tih veza može se vidjeti na Slici 9.



Slika 9: Hipoteza o epizodičkom pamćenju (Jiang i Forster, 2001). J1 predstavlja leksičke unose za riječ na dominantnom jeziku, a J1* epizodički zapis svih prijašnjih iskustava s tom riječi. J2* predstavlja epizodički zapis prijevodnoga ekvivalenta na nedominantnom jeziku. Riječi su na J2 predstavljeni jedino kao epizodički zapisi. Odluke epizodičkoga prepoznavanja temelje se na podacima epizodičke memorije, a leksičke se odluke temelje na podacima leksičke memorije. Prilagođeno prema Jiang i Forsteru (2001).

U svojim su eksperimentima Jiang i Forster (2001) s kinesko-engleskim dvojezičnim govornicima potvrdili da u zadatku epizodičkoga prepoznavanja (*episodic recognition task*)³ i u zadatku leksičke odluke do učinka usmjeravanja može samo doći kada su usmjerivač i ciljna riječ prethodno prisutni u epizodičkom pamćenju. Isto tako u zadatku leksičke odluke do učinka usmjeravanja može doći samo u smjeru J1-J2.

Da bi ispitanik mogao donijeti leksičku odluku, mora aktivirati leksičko pamćenje. Kako bi se uspostavila veza između J2* i J1, ispitanik treba svjesno zamijetiti i obraditi usmjerivač na J2. Budući da su epizodičko i leksičko pamćenje odvojeni, za aktivaciju veze koja prelazi iz jednoga pamćenja u drugo treba postojati nadređen sustav kontrole koji se može aktivirati jedino ako se svjesno zamijeti J2*. Do toga ne može doći ako se u eksperimentu upotrebljavaju maske.

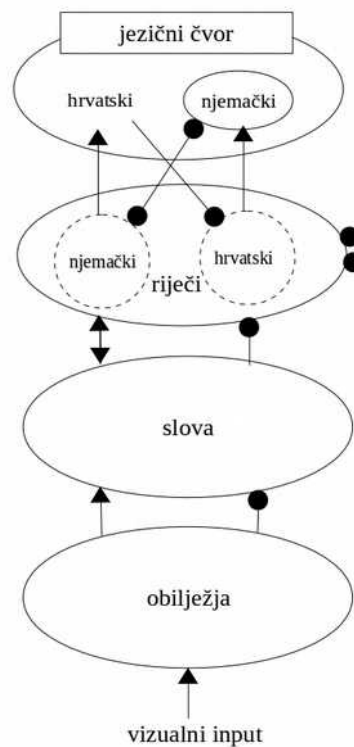
Kako u zadatku epizodičkoga pamćenja ne dolazi do aktivacije leksičkoga pamćenja, učinak se usmjeravanja pojavljuje samo u smjeru J2-J1. Tu za Jianga i Forstera (2001) iznenađujuću pojavu objašnjavaju time što u smjeru J1-J2 riječi na J1 nisu bile prisutne u fazi učenja koja je prethodila samom eksperimentu. Naime, ne može doći do usmjeravanja između usmjerivača i ciljne riječi ako je jedna smještena u epizodičkom, a druga u leksičkom pamćenju. Witzel i Forster (2012) su, međutim, uspjeli u eksperimentu sa zadatkom epizodičkoga prepoznavanja dokazati učinak usmjeravanja u oba smjera, J2-J1 i J1-J2. Iz toga proizlazi, kako Witzel i Forster tvrde, da podjela prema kojoj je J1 smješten u leksičkom, a J2 u epizodičkom pamćenju više nije održiva. Usmjerivači na J1 aktiviraju svoju leksičku reprezentaciju uvijek i neovisno o zadatku. Zbog toga što je ciljna riječ na J2 prisutna u epizodičkom pamćenju, dolazi do usmjeravanja.

8.5. Bilingvalni interaktivni aktivacijski model plus (BIA+)

Bilingvalni interaktivni aktivacijski model plus (Dijkstra i van Heuven, 2002) nastao je na temelju bilingvalnoga interaktivnog aktivacijskog modela (Dijkstra i van Heuven, 1998; Dijkstra, Grainger i van Heuven, 1999). Sam se model BIA zasniva na jednojezičnom modelu interaktivne aktivacije (McClelland i Rumelhart, 1981). To je lokalistički konektivistički računalni model vidnoga prepoznavanja riječi.

³ U zadatku epizodičkoga pamćenja ispitanicima se prije nego što počne ispitivanje daje popis riječi koji trebaju naučiti. Tijekom ispitivanja trebaju odlučiti je li se pojedina riječ nalazila na popisu koji su prije ispitivanja proučavali ili ne.

Ključna je zadaća modela BIA prepoznavanje ortografskih reprezentacija kod dvojezičnih govornika. BIA se sastoji od četiri razine: obilježja, slova, riječi i jezika. Svaki je element predstavljen čvorom. Oni se međusobno mogu pobuditi ili suspregnuti u oba smjera odozgo prema dolje i odozdo prema gore. Niz se slova koji može predstaviti riječ u modelu BIA obrađuje tako da se pobuđuju vidna obilježja za svako pojedino slovo predstavljenoga niza slova. Ta obilježja dalje pobuđuju slova koja ih sadržavaju. Potom se pobuđivanje širi na sljedeću višu razinu – razinu riječi. Ondje se pobuđuju riječi iz oba jezika pri čemu pobuđeno slovo mora biti na odgovarajućem mjestu u riječi. Istodobno se sve ostale riječi potiskuju. Prema tome, na razini se riječi sve riječi međusobno potiskuju bez obzira na jezik kojemu pripadaju. Pobuđene pak riječi pobuđuju dalje čvorove na razini jezika koji potom potiskuju sve riječi iz drugoga jezika (za detaljan prikaz modela BIA vidi Sliku 10).



Slika 10: Bilingvalni interaktivni aktivacijski model (BIA) (Dijkstra i van Heuven, 1998). Sastoji se od četiri razine: obilježja slova, slova, riječi i jezika. Prepoznavanjem pojedinih obilježja slova, pobuđuju se slova koja imaju ta obilježja te istodobno potiskuju ona koja ih nemaju. U drugom koraku aktivacija prelazi s razine slova na razinu riječi koje se sastoje od tih slova na odgovarajućim mjestima. Sve se ostale riječi uslijed toga potiskuju bez obzira na to kojem jeziku pripadaju. Riječi u nastavku pobuđuju čvor pripadajućega jezika koji potom potiskuje sve riječi drugoga jezika. Strelice s trokutastom glavom označavaju veze koje pobuđuju čvorove na sljedećoj ili istoj razini, a strelice s okruglom glavom veze koje ih potiskuju. Prilagođeno prema Dijkstra i van Heuvenu (1998).

Pobuđivanje je dvojako: ili postoji ili ne postoji. Ako postoji, ono se širi dalje na čvorove sljedeće razine. Pobuditi se mogu čvorovi iste i susjedne razine. Tako, primjerice, obilježja pojedinih slova aktiviraju čvorove na sljedećoj, višoj razini odozdo prema gore, dok istodobno čvorovi na razini jezika potiskuju aktivaciju čvorova odozgo prema dolje tako da koče aktivaciju riječi iz drugoga jezika. Osim toga potiskivanje djeluje i na susjedne čvorove iste razine tako što se, primjerice, potiskuju riječi kojima nisu aktivirana odgovarajuća slova. Isto tako riječi mogu utjecati na slova odozgo prema dolje potvrđujući ona od kojih su sazdane. Koja će se pak riječ najbrže prepoznati, ovisi o pragu prepoznavanja (*recognition threshold*). Drugim riječima, čim pobuđeni čvorovi riječi prenesu svoju aktivaciju na jezični čvor, on potiskuje sve riječi iz drugoga jezika, a istodobno se riječi međusobno potiskuju sve do onoga trenutka kada se dosegne prag prepoznavanja. Stoga se može reći da vrijeme reakcije u biti odgovara vremenu dosezanja praga prepoznavanja. Valja napomenuti da temeljna aktivacija (*baseline level of activation*) nije jednaka za sve čvorove na razini riječi (de Groot, 2011). Ona ovisi, primjerice, o tome kada je zadnji put upotrijebljena riječ, ali i o frekvenciji upotrebe riječi. Riječi koje se češće upotrebljavaju imaju višu temeljnu aktivaciju. Tako se frekventnije riječi prepoznaju brže od manje frekventnih riječi.

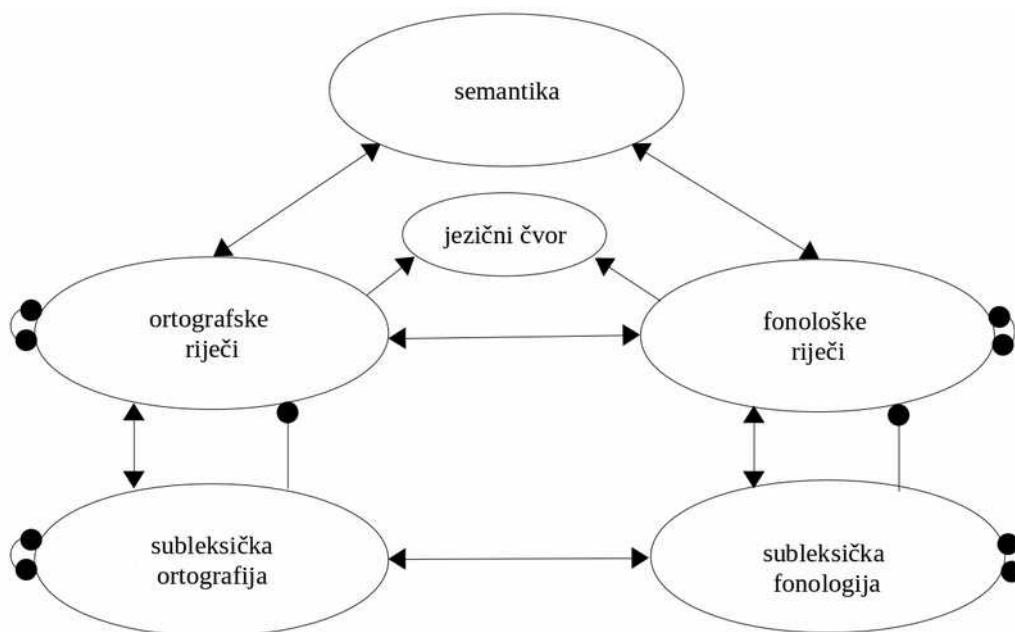
Ključnu ulogu, prema tome, u modelu BIA igraju čvorovi. Oni imaju dvojaku funkciju. Djeluju kao jezične reprezentacije i kao nelingvistički mehanizmi. Prvo se očituje tako što Dijkstra i van Heuven (1998) smatraju da jezični čvorovi služe kao jezične oznake koje upućuju na pripadnost određenom jeziku ili mogu akumulirati aktivaciju tako da utječu na prepoznavanje riječi u daljnjem tijeku eksperimenta. Kao nelingvistički mehanizmi djeluju kada utječu na buduću aktivaciju konteksta usmjeravajući tako pozornost i očekivanja ispitanika. Na taj način čvorovi ubrzavaju prepoznavanje riječi. Kako BIA polazi od hipoteze o integriranom leksikonu, na prepoznavanje ciljne riječi utječu i susjedi iz ciljnoga i neciljnoga jezika. Isto tako BIA polazi od neselektivnoga pristupa leksikonu.

S vremenom se pokazalo da BIA ima stanovita ograničenja koja su Dijkstra i van Heuven (2002) pregledno sveli na sljedeće točke:

1. ne postoji fonološka i semantička reprezentacija
2. uloga međujezičnih homografa i kognata nije dovoljno jasno precizirana
3. jezični su čvorovi nejasno određeni
4. vrlo je ograničen način utjecaja jezičnih i nejezičnih konteksta na prepoznavanje riječi
5. ne postoji detaljan opis kako ispitanici rješavaju pojedine zadatke

6. nedovoljno je jasno opisan odnos između sustava prepoznavanja riječi i odrednica zadataka

Stoga su Dijkstra i van Heuven nadogradili svoj model tako što su uzeli u obzir spomenute nedostatke. Njihov novi model BIA+ uključuje prijašnji model BIA te bitne odrednice iz Greenova IC modela (Green, 1998). Međutim, prije nego što su izgradili BIA+, napravili su model koji je uključivao semantiku, ortografiju i fonologiju, a nazvali su ga SOPHIA (*Semantic, Orthographic, and Phonological Interactive Activation Model* - semantički, ortografski i fonološki interaktivni aktivacijski model) (de Groot, 2011; Thomas i van Heuven, 2005). Taj model detaljnije opisuje ortografsku reprezentaciju od BIA-je. Između razine slova i riječi postavljene su dodatne dvije razine: razina ortografskih skupova (*orthographic cluster*) i razina ortografskih slogova (*orthographic syllables*). Važno je napomenuti da SOPHIA, za razliku od BIA-je, ne predviđa potiskivanje koje se širi s jezičnih čvorova na čvorove riječi (de Groot, 2011) (vidi Sliku 11).



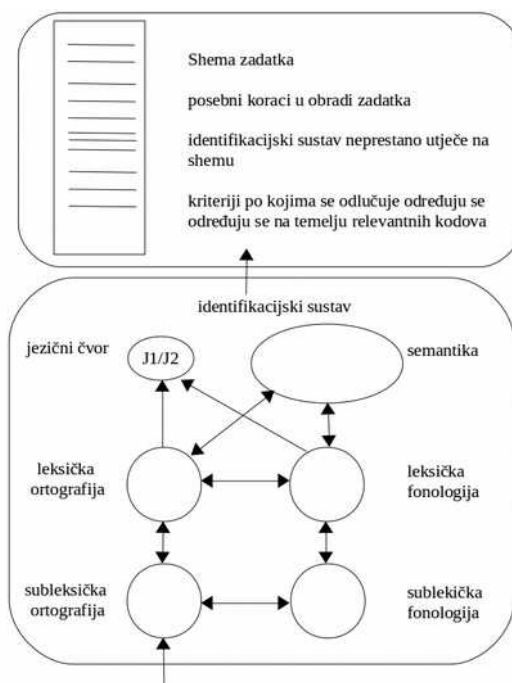
Slika 11: Model semantičke, ortografske i fonološke interaktivne aktivacije (SOPHIA). Strelice s trokutastim vrhom označuju pobuđivačke veze, a strelice s okruglim vrhom označuju inhibicijske veze. Prilagođeno prema de Groot (2011).

BIA+ proširuje pretpostavke koje ima BIA o ortografskim reprezentacijama na fonološke i semantičke. Polazi od toga da ortografska obilježja pobuđuju fonološke reprezentacije te se na taj način širi aktivacija. Međutim, činjenica kako pobuđivanje ovisi o subjektivnoj frekvenciji upućuje na to da će prepoznavanje fonoloških i semantičkih reprezentacija na J2 kasniti za onima na J1 što su van Heuven i Dijkstra nazvali pretpostavkom o privremenom kašnjenju (*temporal delay assumption*). Posljedica će takve

pojave biti dvojaka: (i) međujezični će učinci biti veći u smjeru od J1 prema J2 nego obrnuto i (ii) odsustvo međujezičnih fonoloških i semantičkih učinaka može se pojaviti ukoliko postavke zadatka omogućuju bržu reakciju na ortografske i slične kodove.

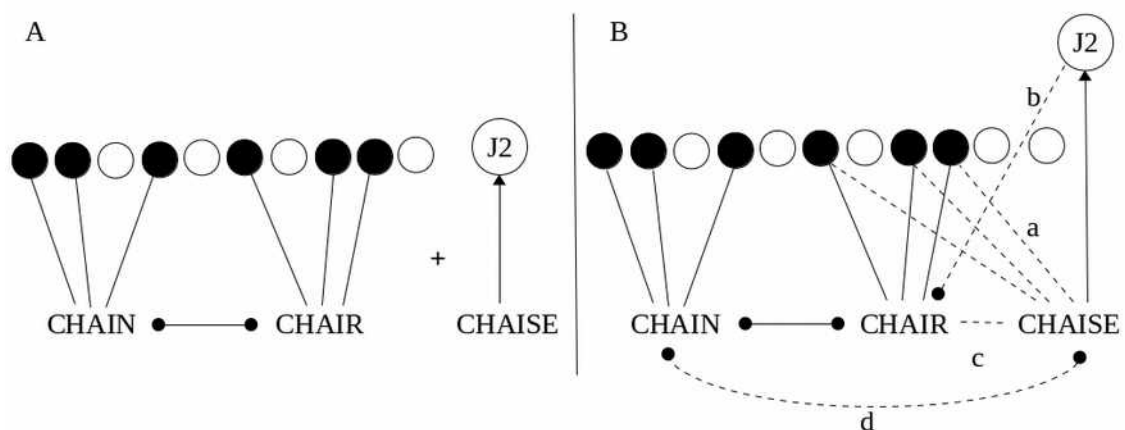
Dijkstra i van Heuven pristupili su u svojem modelu BIA+ sasvim novom određivanju jezičnih čvorova koji su prije bili nejasno definirani. Ograničili su im funkciju. Više ne služe kao jezični filter koji će pobuđivati ili potiskivati određene riječi niti će služiti kao mehanizam koji unaprijed pobuđuje informacije dobivene iz konteksta i na taj način utječe na obradu riječi. Međutim, zadržavaju jezične čvorove kao koristan konstrukt u modelu.

U modelu BIA+ postoji razlika između sustava identifikacije riječi i sustava zadatka. Do te su spoznaje došli oslanjajući se na Greena (1998) (za shematski prikaz modela BIA+ vidi Sliku 12). Jezični kontekst može utjecati na sustav prepoznavanja riječi, a nejezični na sustav zadatka. Jezični se pak kontekst definira u smislu učinaka koji se temelje na leksičkim, sintaktičkim i semantičkim pojavama, a nejezični kontekstni učinci nastaju na temelju instrukcija, zahtjeva koje zadatak postavlja pred ispitanika i na osnovi njegovih očekivanja (Dijkstra i van Heuven, 2002). To znači da, primjerice, fonološka, sintaktička i semantička obilježja mogu utjecati na prepoznavanje riječi na J2 na isti način kako to čine i na J1. Očekivanja koja pojedinac ima od zadatka postavljena su na početku eksperimenta, dok se ispitanika upoznaje s time kako će se odvijati eksperiment.



Slika 12: Bilingvalni interaktivni aktivacijski model plus (BIA+) (Dijkstra i van Heuven, 2002). Sustav prepoznavanja riječi proširen je s isključivo ortografskih reprezentacija riječi na fonološke i semantičke reprezentacije. Potiskivajuće veze koje su postojale u BIA-ji, u BIA+ su ispuštene. Izvanjezični kontekst može utjecati na izvršavanje zadataka. To je obuhvaćeno shemom jezičnoga zadatka koja je uvedena pod utjecajem Greenova IC modela (Green, 1998). Prilagođeno prema Dijkstra i van Heuvenu (2002).

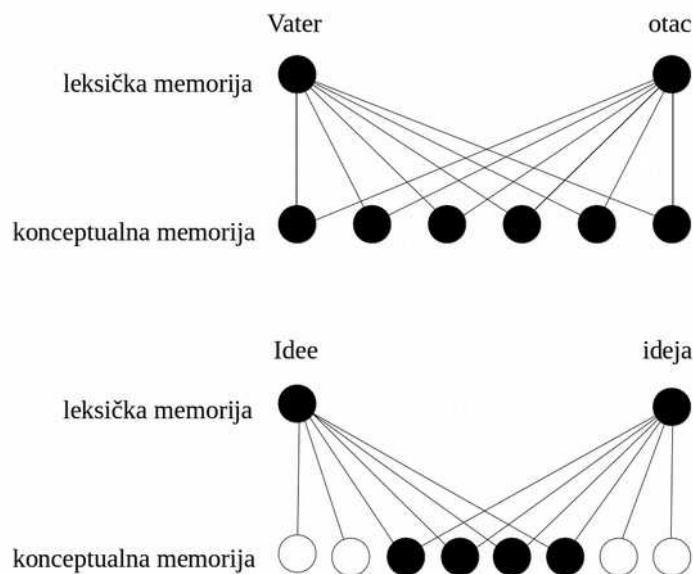
Grainger, Midgley i Holcomb (2010) iznijeli su razvojni model BIA (*BIA developmental*, BIA-d) prikazan na Slici 13. BIA-d se temelji na usvajanju jezika i na ograničenjima koja proizlaze iz učenja riječi na J2. Polazi se od dvojezičnih govornika čiji je J1 već razvijen. Prema tome, model se BIA-d odnosi na početne dvojezične govornike koji J2 uče u razrednom okruženju. Isto se tako taj model odnosi na dvojezične govornike čiji J1 i J2 dijele isto pismo. Svaka je riječ povezana s više različitih semantičkih obilježja. Ako određene riječi aktiviraju isti podražaj (ortografski slična riječ) i ako one pri tome ne dijele semantička obilježja, te su međusobne veze inhibicijske. Prema tom modelu početno je učenje nadzirano tako da se svakoga učenika J2 upućuje na prijevod neke riječi iz J2 na J1. Na taj se način aktivira riječ na J2, ali i pripadajuća riječ na J1 te njezina semantička reprezentacija. Kada riječ na J2 dosegne određenu razinu kritičke aktivacije, a da se pri tome ne povezuje sa svojim značenjem na J1, J2 postaje autonoman u odnosu na J1. Autonomija se pojačava inhibicijom odozgo prema dolje od jezičnoga čvora J2 do prijevodnoga ekvivalenta na J1.



Slika 13: Razvojni model bilingvalne interaktivne aktivacije (BIA-d) (Grainger, Midgley i Holcomb, 2010). Inhibicijske su veze označene krugovima, a ostale su veze pobuđivačke. Isprekidane linije označavaju one veze kod kojih dolazi do najveće promjene tijekom učenja J2. Dio A prikazuje početno učenje francuskoga J2 kod dvojezičnoga govornika kojem je engleski J1. Riječi na J1 (chain i chair) međusobno su povezane inhibicijskim vezama, a pobuđivačkim su vezama povezane sa semantičkim obilježjima. Govorniku je prikazana riječ na J2 (chaise). Dio B prikazuje razvojne promijene u vezama između forme i značenja riječi kao funkcije izloženosti J2. Veze označene s a označavaju postepeno pojačavanje pobuđivačkih veza između oblika riječi na J2 prema semantičkim obilježjima. Veze označene s b označavaju postepeno povećanje inhibicijskih veza od jezičnoga čvora J2 prema riječima na J1. Veze označene s c označavaju pojavu kako se postepeno povećava pobuđivanje riječi na J2 i njihovih prijevodnih ekvivalenata na J1. Kako se povećava inhibicija od jezičnoga čvora J2, smanjuje se pobuđivanje tih veza. Istodobno se odbacuje povezivanje riječi na J2 s riječima na J1 (clamping). Veze označene s d označavaju proces razvoja inhibicijskih veza od riječi na J2 prema drugim riječima na J2 i J1 koje su ortografski slične riječima na J2. Preuzeto od Graingera, Midgley i Holcoma (2010).

8.6. Model distribucijskih konceptualnih obilježja (DCFM)

Model distribucijskih konceptualnih obilježja (*Distributed conceptual feature model* - DCFM) izložila je de Groot (1992a; 1992b; 1993). Taj model predviđa da ispitanici brže prevode konkretne riječi nego apstraktne. Osim toga konkretne riječi pokazuju veću semantičku sličnost na J1 i J2 od apstraktnih. To je povezano s mjerom veće konkretnosti i predočivosti konkretnih riječi. De Groot smatra da se konkretne riječi razlikuju od apstraktnih po tome koliko konceptualnih čvorova dijele riječi na J1 i J2. Drugima riječima, razlikuju se po stupnju preklapanja prijevodnih ekvivalenata. Primjerice, konkretne riječi *Vater* i *otac* dijele sve svoje konceptualne reprezentacije, dok riječi *Idee* i *ideja* samo neke (vidi Sliku 14). Što se više konceptualnih elemenata dijeli između prijevodnih ekvivalenata, to će se brže širiti aktivacije od leksičkih čvorova jedne riječi prema njezinu prijevodu.



Slika 14: Model distribucijskih konceptualnih obilježja (DCFM) (de Groot, 1992a; 1992b; 1993). U gornjem primjeru konkretne riječi *Vater* i *otac* dijele sve svoje konceptualne reprezentacije. U donjem primjeru apstraktne riječi *Idee* i *ideja* dijele samo neke od svojih konceptualnih reprezentacija. Prilagođeno prema de Groot (1992a).

Međutim, prijevodi se ne preklapaju uvijek u svojim semantičkim obilježjima između dvaju jezika. Česta je pojava da riječ na J1 može imati više različitih prijevoda na J2. Basnight-Brown i Altarriba (2015) ističu da se španjolska riječ *ladrón* može prevesti kao *robber*, *thief* ili *burglar* na engleski. Stoga riječi koje imaju više značenja i prijevoda pokazuju veće preklapanje unutar i između jezika. Zbog toga bi se trebali i brže prepoznati (Basnight-Brown i Altarriba, 2015).

Basnight-Brown (2014) identificirala je nekoliko nedostataka u ovom modelu. DCFM ne daje nikakva specifična predviđanja o tome kako se preklapanje obilježja može odraziti na prepoznavanje pisane i izgovorene riječi. Osim toga nije ništa rečeno o tome kakve se informacije kriju iza konceptualnih obilježja.

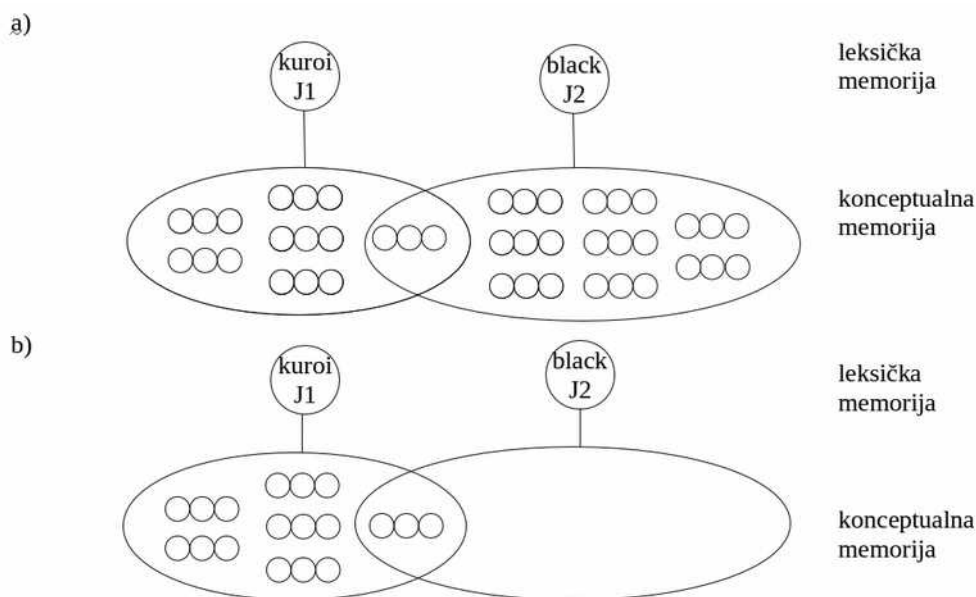
8.7. Model značenja

Model značenja (*Sense Model*) distribucijski je model koji su predstavili Finkbeiner i sur. (2004) kao reakciju na model distribucijskih konceptualnih obilježja (*Distributed conceptual feature model* – DCFM) (de Groot, 1992a; de Groot, 1992b; de Groot, 1993).

Jedan je od ključnih nedostataka DCFM-a taj što ne daje objašnjenje zašto postoji asimetrija u usmjeravanju između dvaju jezika. Prema tom modelu ne bi trebala postojati razlika između usmjeravanja u smjeru J1-J2 i J2-J1 jer se u oba slučaja aktiviraju ista konceptualna obilježja (Finkbeiner i sur., 2004). Kako bi se ta referencijalna asimetrija objasnila, Finkbeiner i sur. uvode svoj model značenja.

Budući da su dvojezični govornici vrsniji na svojem J1 nego na J2, riječi na J1 imat će više značenja nego riječi na J2. Znat će, dakle, manje značenja riječi na J2 nego na J1. To pak znači da će u usmjeravanju u smjeru J1-J2 omjer usmjerenih značenja na J2 biti vrlo velik jer su gotovo sva značenja riječi na J2 povezana s odgovarajućim značenjima na J1. No u obrnutom će smjeru usmjeravanje biti slabije jer postoji mnogo riječi na J2 koje ne dijele ista značenja s J1, te se neće aktivirati dovoljan broj značenja kako bi došlo do usmjeravanja (Finkbeiner i sur., 2004; Wang i Forster, 2010; Xia i Andrews, 2015) (vidi Sliku 15).

Model značenja ima različite učinke usmjeravanja u zadatku leksičke odluke i u zadatku semantičke kategorizacije kako su to u svojem istraživanju utvrdili Finkbeiner i sur. (2004). Zaključili su kako se u zadatku leksičke odluke s maskiranjem pojavljuje učinak usmjeravanja u smjeru J1-J2, no ne i u obrnutom smjeru J2-J1. Isto su tako potvrdili da se u zadatku semantičke kategorizacije učinak usmjeravanja pojavljuje u oba smjera J1-J2 i J2-J1. Ustanovili su da važnu ulogu u tome igra kategorija jer djeluje kao filter koji regulira aktivaciju samo onih značenja koja su relevantna za ciljnu riječ. Kako u zadatku leksičke odluke ne postoji unaprijed zadana kategorija, ne postoji ni filter koji bi poboljšao učinak usmjeravanja (Finkbeiner i sur., 2004; Wang i Forster, 2010).



Slika 15: Model značenja reprezentacija dvojezične memorije (Finkbeiner i sur., 2004). Svaki niz od triju krugova predstavlja jedno značenje japanske riječi kuroi ili engleske riječi black. Obje riječi, kuroi i black, dijele jedno značenje, a to je značenje boje. Svaka riječ također posjeduje više nizova značenja koji su specifični za pojedini jezik. (a) predstavlja memorijsku reprezentaciju riječi kuroi i black kod japansko-engleskih dvojezičnih govornika visoke razine vrsnosti. (b) predstavlja memorijsku reprezentaciju riječi kuroi i black kod govornika koji su usvojili samo značenje boje black, ali ne i druga specifična značenja riječi black koja ta riječ ima na engleskom jeziku. Prilagođeno prema de Groot (2011).

Xia i Andrews (2015) osporavaju neke od zaključaka do kojih su došli Finkbeiner i sur. (2004) jer su nastali na temelju istraživanja s jednojezičnim ispitanicima. Finkbeiner i sur. tvrde da riječi s više značenja na J1 odgovaraju riječima na J1, a riječi s manje značenja na J1 odgovaraju riječima na J2. Tako su otkrili da ne postoji učinak usmjeravanja u zadatku leksičke odluke gdje su umjerivači bili riječi s malo značenja (analogne riječima na J2), a ciljne riječi one s mnogo značenja (analogne riječima na J1). No u obrnutom se smjeru pojavio učinak usmjeravanja. U zadatku semantičke kategorizacije pak postoji simetrija u oba smjera. Xia i Andrews (2015) u svojem su istraživanju s dvojezičnim ispitanicima samo djelomično mogli potvrditi rezultate do kojih su došli Finkbeiner i sur. (2004) te Jiang i Forster (2010). Zaključili su da semantičko usmjeravanje u zadatku semantičke kategorizacije nije simetrično, već je isto tako asimetrično kao i u zadatku leksičke odluke samo u manjem omjeru. Kategorija koja se kod Finkbeinera i sur. (2004) smatrala filtrom nije uklonila asimetriju, već ju je samo smanjila.

9. CILJ ISTRAŽIVANJA

Cilj je ovoga istraživanja otkriti može li računalni model *word2vec* u svojoj izvedbi CBOW pomoći u predviđanju i zaključivanju o tome kako funkcionira mentalni leksikon dvojezičnoga govornika. Želi se otkriti može li model CBOW dati uvid u njegovo funkcioniranje te jesu li učinci usmjeravanja slični onima kod ljudi i što je uzrok tomu.

Ključan problem u istraživanjima koja su se do sada bavila zadacima leksičke odluke s usmjeravanjem jest taj što su se analize rezultata dobivenih od lista usmjerivača i ciljnih riječi koje su oblikovali istraživači vršile DSM-ovima uvijek *post hoc*. Na taj su se način razmatrale samo ona istraživanja koja su pokazala znakovite statističke promjene u usmjeravanju. Ovim istraživanjem želimo, naime, odrediti predviđačku snagu DSM-a kao jednu od varijabli kojom se može manipulirati prilikom konstruiranja eksperimenata. Stoga će se u ovome istraživanju poći suprotnim smjerom: pomoću DSM-a će se generirati lista usmjerivača i zadanim ciljnim riječima. Ti će se parovi riječi potom testirati u zadatku leksičke odluke s usmjeravanjem s govornicima njemačkoga jezika kao J2. Odlučili smo se za tu vrstu zadatka jer pokazuje veće učinke usmjeravanja nego, primjerice, zadatak imenovanja (Brysbaert i sur., 2014; Hutchison, 2003; Lucas, 2000).

9.1. Hipoteze istraživanja

U ovom istraživanju polazimo od sljedećih hipoteza:

H1: Vrijeme reakcije riječi ponuđenih kao najbližnje traženim riječima prema DSM-u *word2vec* odražavat će vrijeme reakcije riječi prikupljene diskretnim slobodnim asocijacijama. To je u skladu s istraživanjem Günthera, Rinaldija i Marellija (2019) koji smatraju da DSM ne smije biti bolji u rješavanju bihevioralnih zadataka od čovjeka.

H2: Ispitanici će u zadatku leksičke odluke s usmjeravanjem imati najbrže reakcije kod onih riječi koje su najbližnje usmjerivaču. To je u skladu s modelom BIA+. Prema tom će se modelu riječi predstavljene čvorovima nalaziti na kraćoj međusobnoj udaljenosti i zbog toga će se brže aktivirati.

H3: Vrijeme reakcije će biti kraće za njemačke riječi ukoliko im je prethodio usmjerivač na hrvatskom. To je u skladu s modelom RHM. Budući da dvojezični govornici imaju bogatiju semantičku reprezentaciju riječi na J1 nego na J2, brže će se aktivirati riječi na J1 nego na J2.

10. EKSPERIMENT 1: ZADATAK LEKSIČKE ODLUKE S USMJERAVANJEM S DISKRETNIM SLOBODNIM ASOCIJACIJAMA

10.1. Metoda

10.1.1. Ispitanici

U eksperimentu sudjelovalo je sveukupno 54 ispitanika (41 ženskih i 13 muških) koji su to učinili dragovoljno i anonimno te su mogli odustati u svakom trenutku. Ispitanici su bili različite dobi ($M = 35.481$ godina, $SD = 6.667$). Najmlađi je ispitanik bio 17, a najstarija ispitanica 50 godina. Ispitanici odudaraju od tipičnoga uzorka ispitanika koji sudjeluju u psiholingvističkim istraživanjima. Uobičajeno je da u takvim istraživanjima sudjeluje studentska populacija jer je ona istraživačima najdostupnija. Henry (2008) ističe da je ta populacija istraživačima najpovoljnija u pogledu novca i vremena. Osim toga teško je prikupiti dovoljan broj ispitanika koji su voljni odvojiti vrijeme za takva istraživanja. S druge pak strane, takvi ispitanici daju drugačiju sliku na fenomene koji se istražuju u odnosu na one koji nisu studenti (Henry, 2008; Sears, 1986). Stoga smo se u ovom radu odlučili za drugačiju strukturu ispitanika. Drugi je razlog bio taj što je u jeku pandemije nove bolesti dišnih puteva COVID-19 pristup školama, fakultetima i drugim uobičajenim mjestima na kojima se izvode istraživanja bio onemogućen. Mnogi su ispitanici potvrdili u osobnom razgovoru s istraživačem da se prvi puta susreću s takvim istraživanjem. Međutim, nisu pokazivali nikakve poteškoće u rješavanju zadataka. Ispitanici su uglavnom prijatelji, kolege i znanci istraživača.

Većina je ispitanika visoke stručne spreme i u svakodnevnom je kontaktu s njemačkim jezikom iz profesionalnih ili privatnih razloga jer su nastavnici i učitelji njemačkoga jezika u javnim i privatnim školama i učilištima ili rade u tvrtkama koje su orijentirane na njemačko govorno područje ili dolaze s toga područja. Većini ($N = 32$) njemački je jezik bio prvi strani jezik, a ostatku ($N = 22$) prvi je strani jezik bio engleski. Upitnikom koji su ispitanici ispunili nakon glavnoga dijela eksperimenta ustanovili smo da nema signifikantne razlike u samoprocjeni vještina čitanja ($t(42.898) = .345, p > .05$), govorenja ($t(44.661) = .995, p > .05$) i slušnoga razumijevanja ($t(41.167) = 1.201, p > .05$) ako je njemački prvi ili drugi strani jezik. Ispitanici su njemački počeli učiti u prosječnoj dobi od 8.778 godina ($SD = 3.489$). Najviše ga se počeo učiti u školi (61.11%), u nešto manjem broju u stranoj zemlji (22.22%), a

ostatak (16.67%) pomoću televizijskih serija i filmova, internetskih sadržaja, u obitelji i sl. Isto tako nema značajne razlike s obzirom na godine učenja njemačkoga jezika ($t(47.795) = 1.171, p > .05$). Najviše je doprinijelo njihovu poznavanju njemačkoga jezika čitanje na njemačkom jeziku, samostalno učenje, gledanje televizijskih emisija i filmova na njemačkom jeziku te internetski sadržaji na njemačkom jeziku, slušanje radija i glazbe na njemačkom jeziku, interakcija s prijateljima te interakcija s obitelji. Najviše su izloženi njemačkom jeziku kada čitaju, samostalno uče jezik, gledaju televizijske serije i filmove te preko sadržaja na internetskim stranicama. Najmanje su pak izloženi radijskim emisijama i glazbi na njemačkom jeziku te interakciji s prijateljima i obitelji. Unatoč tome što se u Hrvatskoj engleski jezik može smatrati drugim jezikom (Mihaljević Djigunović i Geld, 2003), nema značajne razlike između dnevne izloženosti njemačkom i engleskom jeziku ($t(53) = 1.735, p > .05$). Svi se sudionici koriste redovito njemačkim i hrvatskim jezikom, dobroga su vida ili im je korigiran. Nitko od ispitanika nema jezične ili neurološke poremećaje.

10.1.2. Instrumenti i materijal

Eksperiment je napravljen pomoću programa *OpenSesame* (Mathôt, Schreij i Theeuwes, 2012), a kao pomoćni sloj koji upravlja zaslonom i tipkovnicom (*backend*) odabran je *PsychoPy* (Peirce i sur, 2019). Njime se mjerila brzina odgovora i njihova točnost.

Za ovaj eksperiment konstruirane su dvije liste riječi. Svaka se lista sastojala od 80 parova usmjerivača i ciljnih riječi. Polovicu parova činili su parovi u kojima je ciljna riječ bila pseudoriječ. Pseudoriječi udovoljavale su fonotaktičkim i pravopisnim pravilima hrvatskoga odnosno njemačkoga jezika. Izrađene su pomoću programa *Wuggy* (Keuleers i Brysbaert, 2010). Drugu polovicu činili su parovi u kojima je svakoj ciljnoj riječi određen usmjerivač koji je asocijativno-semantički povezan s ciljnom riječi i usmjerivač koji to nije (Gulan, 2016; McNemara, 2005). Prvi korak u izradi asocijativno-semantički povezanih riječi bio je utvrđivanje koncepta. Konceptima je trebala odgovarati konkretna imenica (vidi Sliku 16). Zatim se odredila riječ koja najviše odgovara tom konceptu u hrvatskom odnosno njemačkom jeziku. U našem je to primjeru *kuća* na hrvatskom i *Haus* na njemačkom jeziku. Ako se hrvatska riječ uzela kao ciljna riječ, njemačkoj je riječi određena asocijativno-semantički povezana (*Dach*) odnosno nepovezana riječ (*Brief*). Te su riječi uzete kao usmjerivači.



Slika 16: Prikaz koncepta. Ovom konceptu u hrvatskom jeziku odgovara riječ kuća, a u njemačkom jeziku riječ Haus. Ako je hrvatska riječ uzeta kao ciljna riječ, njemačka je riječ poslužila kao podražaj u određivanju asocijacija. Preuzeto od Snodgrassa i Vanderwart (1980).

Asocijativno-semantički parovi određeni su pomoću De Deynea i sur. (2018) za njemačke usmjerivače. Projekt *Small World of Words* (De Deyne i sur., 2018) prikuplja primarne, sekundarne i tercijarne asocijacije na podražajne riječi. Podacima se može pristupiti preko internetskoga sučelja www.smallworldofwords.org. Riječi je moguće pretražiti na dva načina. Prvo, mogu se potražiti najčešće asocijacije na određen podražaj, i drugo, najčešći podražaji određenom odgovoru ili asocijaciji. Drugim riječima, mogu se pretražiti asocijacije unaprijed i unazad (*forward* i *backward association*).

Hrvatski su usmjerivači određeni pomoću Vezmar (2017). To je diplomski rad nastao u okviru studija psihologije. Svjesni smo ograničenja takvih radova, no odlučili smo se upotrijebiti materijal iz toga rada jer prema našim saznanjima nije još uspostavljena baza diskretnih slobodnih asocijacija hrvatskoga jezika. Budući da nema prikladne baze, morali smo se njime koristiti. Smatrali smo to boljim nego da se prevode riječi iz njemačke baze diskretnih slobodnih asocijacija. Različite kulturološke i životne okolnosti mogu, naime, dovesti do različitih asocijacija ovisno o tome gdje se pojedinac nalazi i koje je jezične vrsnosti (Racine, Higginbotham i Munby, 2014).

Uvijek se uzimala ona asocijacija koja je najčešća. Međutim, ako dva podražaja imaju istu najčešću asocijaciju ili je asocijacija druga podražajna riječ, uzeta je sljedeća najčešća asocijacija. Isto se nisu uzimali u obzir oni usmjerivači i ciljne riječi koji su međusobni prijevodi.

Nepovezane su se riječi odredile tako da budu što udaljenije od koncepta. Udaljenost između povezanih i nepovezanih riječi odredili smo u programu *R* (The R Core Team, 2018) pomoću programskoga paketa *LSAfun* te pomoću semantičkoga prostora *dewac_cbow* za

njemačke riječi i *hr_cbow* za hrvatske riječi (Günther, Dudschig i Kaup, 2015). Primjerice, za koncept MAJKA (*Mutter*) odredili smo da nepovezana riječ bude *Rad* jer među njima kosinusna sličnost iznosi .072 za razliku od povezane riječi *Vater* gdje ona iznosi .814. Ne postoji signifikantna razlika ni u hrvatsko-njemačkim relevantnim parovima ($t(39) = 0$, $p = 1$) ni u njemačko-hrvatskim relevantnim parovima ($t(39) = -.196$, $p = .846$). Primjere ciljnih riječi i usmjerivača kao i prosječan broj slova po riječi mogu se vidjeti u Tablici 7 i Tablici 8.

Tablica 7: *Primjeri parova usmjerivača i ciljne riječi u eksperimentu s diskretnim asocijacijama. Ciljnoj je riječi uvijek prethodio usmjerivač koji je mogao biti diskretna asocijacija ili semantički i asocijativno nepovezana riječ. Jedna polovica parova koju su vidjeli ispitanici bila je asocijativno-semantički povezana, dok druga polovica nije. Svaki je ispitanik vidio svaku ciljnu riječ samo jednom.*

Njemačko-hrvatski parovi		
ciljna riječ	usmjerivač (diskretna asocijacija)	usmjerivač (nepovezana riječ)
Schaf Wolke Sand Ohr	vuna kiša plaža sluh	porotnik kopito kos govedo
Hrvatsko-njemački parovi		
ciljna riječ	umjerivač (diskretna asocijacija)	usmjerivač (nepovezana riječ)
juha jaje jastuk polica	Löffel Huhn Bett Buch	Kunde Leber Ball Kalb

Tablica 8: *Prosječan broj slova po riječima u eksperimentu s diskretnim asocijacijama. U zagradi su standardne devijacije.*

Njemačko-hrvatski parovi		
ciljna riječ	usmjerivač (diskretna asocijacija)	usmjerivač (nepovezana riječ)
4.7 (1.018)	4.75 (1.056)	5.0 (.906)
Hrvatsko-njemački parovi		
ciljna riječ	usmjerivač (diskretna asocijacija)	usmjerivač (nepovezana riječ)
5.275 (1.467)	5.275 (1.467)	5.825 (1.318)

Omjer riječi i pseudoriječi bio je .50. U svakoj je od listi omjer povezanih i nepovezanih parova bio .25. Jedna polovica parova bila je asocijativno-semantički povezana, dok druga nije. Svaki je ispitanik vidio svaku ciljnu riječ samo jedanput. Usmjerivači su se izmjenjivali. Jezik parova usmjerivača i ciljne riječi također se izmjenjivao među ispitanicima. Od ukupno 54 ispitanika, koliko je sudjelovalo u ovom eksperimentu, 24 je ispitanika prvo vidjelo njemačko-hrvatske parove te hrvatsko-njemačke parove, dok je preostalih 30 ispitanika prvo vidjelo hrvatsko-njemačke parove, a potom njemačko-hrvatske parove. Cjelovit popis ciljnih riječi s asocijativno-semantičkim povezanim i nepovezanim usmjerivačima te pseudoriječima može se naći u prilogima koji prate ovaj rad.

10.1.3. Postupak

Zbog otežane mogućnosti okupljanja i nemogućnosti ulaska u škole, fakultete i slične ustanove, ispitanike smo trebali naći na drugačiji način. U prikupljanju podataka istraživaču su pomogli prijatelji. Kako bi se centralno skupili podaci istraživanja, poslužili smo se internetskim alatom *Just Another Tool for Online Studies* – JATOS (Lange, Kühn i Filevich, 2015) na internetskom serveru *Amazon Web Servicesa (AWS)*. Taj alat dodjeljuje zasebnu internetsku adresu svakom ispitaniku pomoću koje pristupa istraživanju koje se potom skida i lokalno čuva na računalu. Kako bi se očuvala anonimnost ispitanika, alat ne čuva njihovu IP-adresu. Po završetku istraživanja podaci se pohranjuju na server odakle ih se može skinuti u *txt*-formatu. Pomoću programa *OpenSesame* (Mathôt, Schreij i Theeuwes, 2012) podaci se pretvaraju u *csv*-format nakon čega ih se može statistički obraditi.

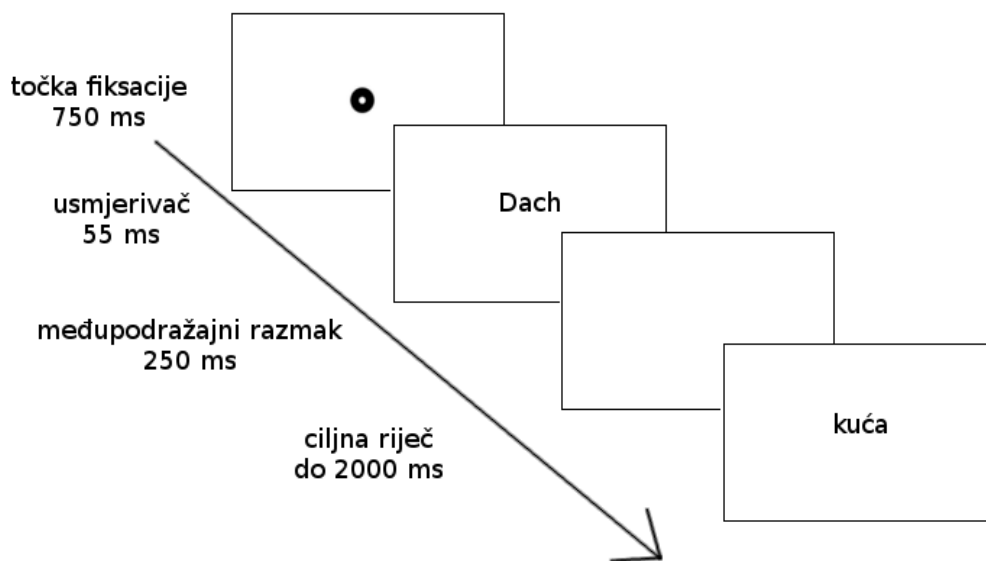
U ovom se istraživanju upotrijebio zadatak leksičke odluke s usmjeravanjem. To znači da je svakoj ciljnoj riječi prethodio usmjerivač. Usmjerivač je mogao biti asocijativno-semantički povezan ili ne. Usmjerivač i ciljna riječ nisu bili na istom jeziku.

Upute tijekom istraživanja bile su napisane hrvatskim jezikom na zaslonu računala. Korištena su crna slova fonta Liberation Sans veličine 30 na bijeloj podlozi. Ispitanici su upoznati s time da se istraživanje sastoji od dva dijela. Prvi je dio bio na hrvatskom, a drugi na njemačkom odnosno obratno. Ispitanicima su dani primjeri za postojeće i nepostojeće riječi na hrvatskom odnosno njemačkom jeziku.

Na početku bi se pojavila točka fiksacije koja je služila tome da odvoji svaki par usmjerivača i ciljne riječi od sljedećega. Bila je vidljiva 750 ms. Nakon točke fiksacije pojavio bi se usmjerivač u trajanju od 55 ms nakon kojega je slijedila ciljna riječ. Ona je bila

vidljiva najviše dvije sekunde. Između usmjerivača i ciljne riječi nalazio se kratak međupodražajni interval od 250 ms za vrijeme kojega je zaslon bio prazan odnosno bijel. Kako nije bilo maskiranja, ispitanici su bili upozoreni na to da će prije riječi na koju trebaju reagirati biti kratko vidljiva riječ na drugom jeziku.

Ukoliko je riječ na zaslonu bila postojeća riječ na hrvatskom ili njemačkom jeziku, ispitanici su trebali pritisnuti tipku "j" na tipkovnici, a ukoliko je riječ na zaslonu bila nepostojeća riječ na jednom od dvaju jezika, trebali su pritisnuti tipku "f". Prije nego što je eksperiment započeo, ispitanici su bili zamoljeni da postave svoj desni kažiprst na tipku "j", a svoj lijevi kažiprst na tipku "f". Čim bi ispitanik pritisnuo jednu od tih dviju tipki, eksperiment bi krenuo i pojavio bi se prvi par riječi. Ispitanici su zamoljeni da reagiraju što točnije i što brže pritiskom na odgovarajuću tipku. Ako su ispitanici pogrešno odgovorili na podražaj, na zaslonu bi vidjeli riječ "netočno" napisanu crvenim slovima. Ondje bi se zadržala 750 ms. Tijek eksperimenta može se vidjeti na Slici 17.



Slika 17: *Tijek eksperimenata.*

Prije relevantnoga dijela eksperimenta ispitanici, koji uglavnom nisu bili upoznati s takvom vrstom zadatka, imali su 20 parova usmjerivača i ciljnih riječi za vježbu. Vježba bi se ponovila ispred svakoga od dvaju relevantnih dijelova eksperimenta. To znači da su imali 20 parova za vježbu ispred hrvatskoga dijela (njemački usmjerivač i hrvatska ciljna riječ) i isto toliko parova ispred njemačkoga dijela (hrvatski usmjerivač i njemačka ciljna riječ). Ispred

svakoga dijela ispitanici bili su upoznati s načinom kako će se izvesti eksperiment. Eksperiment je trajao oko 15 minuta.

Kada su ispitanici bili gotovi sa zadatkom leksičke odluke s usmjeravanjem, podaci bi se pohranili na serveru, a ispitanike bi se istodobno preusmjerilo na drugu internetsku stranicu gdje su ispunjavali kratak upitnik o svojoj jezičnoj vrsnosti i biografiji. Upitnik je nastao na temelju istraživanja koje su provele Marian, Blumenfeld i Kaushanskaya (2007). Njihov upitnik o jezičnom iskustvu i vrsnosti (*LEAP-Q*) pokazao se kao učinkovito i pouzdano sredstvo za procjenu dvojezičnoga statusa pojedinca. To su utvrdile usporedbom samoprocjene vrsnosti dvojezičnih govornika s njihovim jezičnim kompetencijama u oba jezika mjerene objektivnim testovima.

U prilogima koji prate rad može se naći detaljniji pregled sličica s objašnjenjima za ispitanike te upitnik koji su ispunjavali.

10.2. Rezultati

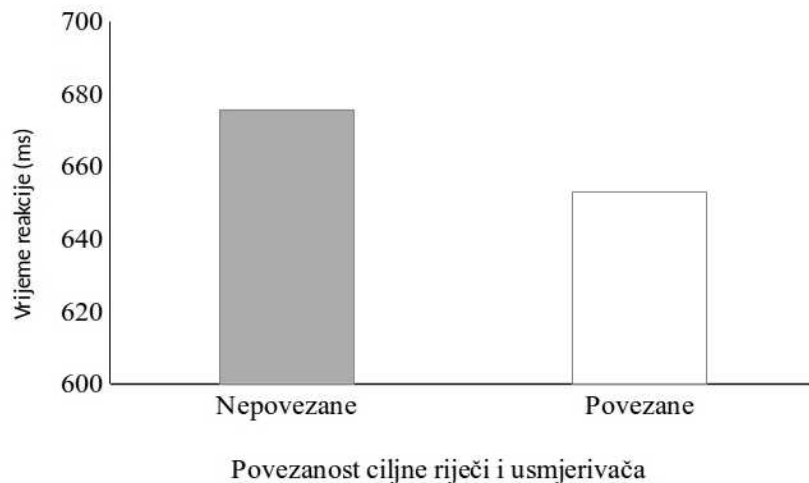
Pri analizama vremena reakcije u oba su eksperimenta u obzir uzeti samo točni odgovori ispitanika. U prvom eksperimentu zabilježeno je ukupno 3.3 % netočnih odgovora koji su isključeni iz daljnjih analiza. Iz analiza su također isključeni odgovori koji su bili kraći od 200 ms (0.19 %) te odgovori duži od 1650 ms (1.04 %). Prije nego je provedena glavna analiza izračunata su prosječna vremena reakcije za svakog od 54 sudionika u svakom od četiri eksperimentalna uvjeta. U Tablici 9 prikazani su rezultati deskriptivne statistike za vrijeme reakcije s obzirom na povezanost usmjerivača i ciljne riječi te s obzirom na jezik na kojem je zadana ciljna riječ.

Tablica 9: Mjere deskriptivne statistike u prvom eksperimentu za vrijeme reakcije s obzirom na povezanost ciljne riječi i usmjerivača te jezik na kojem je zadana ciljna riječ ($N = 54$).

Povezanost riječi	Jezik ciljne riječi	Aritmetička sredina (M)	Standardna devijacija (SD)
nepovezane	hrvatski	639.17	96.30
	njemački	711.92	110.29
povezane	hrvatski	620.22	97.82
	njemački	685.80	108.16

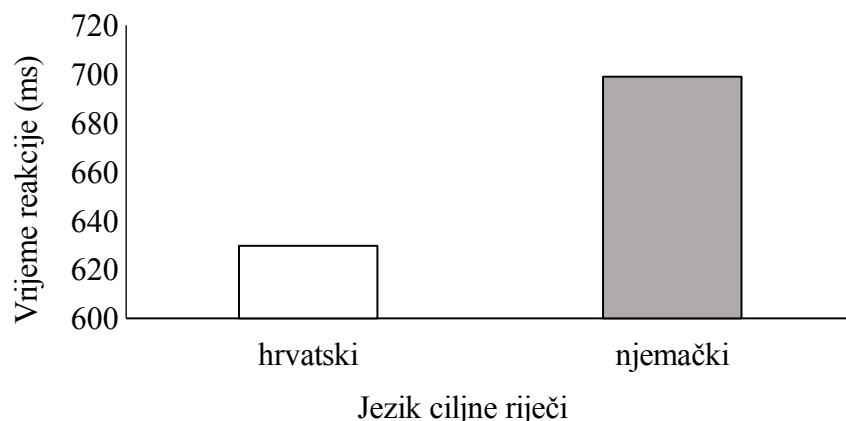
Kako bismo provjerili istraživačke hipoteze, provedena je dvosmjerna analiza varijance s ponovljenim mjerenjima. Interakcija između dva faktora (povezanost riječi \times jezik

ciljne riječi) nije bila statistički značajna. Glavni efekt povezanosti ciljne riječi i usmjerivača je bio statistički značajan: $F(1, 53) = 14.78$; $p < .001$; $\eta^2 = .218$. Pokazalo se da su sudionici prvoga eksperimenta imali brže vrijeme reakcije kada su ciljna riječ i usmjerivač bili asocijativno povezani ($M = 653.01$) nego kada su bili nepovezani ($M = 675.54$). Slika 18 prikazuje glavni efekt povezanosti riječi u prvom eksperimentu.



Slika 18: Prosječne vrijednosti vremena reakcije u prvom eksperimentu s obzirom na povezanost ciljne riječi i usmjerivača.

Glavni efekt jezika na kojem je bila zadana ciljna riječ je također bio statistički značajan: $F(1, 53) = 23.15$; $p < .001$; $\eta^2 = .304$. Sudionici su imali brže vrijeme reakcije kada je ciljna riječ bila na hrvatskom, a usmjerivač na njemačkom jeziku ($M = 629.69$) nego u obrnutoj situaciji ($M = 698.86$). Slika 19 prikazuje glavni efekt jezika ciljne riječi u prvom eksperimentu.



Slika 19: Prosječne vrijednosti vremena reakcije u prvom eksperimentu s obzirom na jezik na kojem je zadana ciljna riječ.

10.3. Rasprava

U prvom smo eksperimentu pokušali ustanoviti postoji li semantičko usmjeravanje između usmjerivača i ciljne riječi koji su na različitim jezicima, hrvatski jezik kao J1 i njemački jezik kao J2. Isto se tako pokušalo odrediti u kojem se smjeru odvija usmjeravanje.

Do sada se u mnogim jednojezičnim istraživanjima pokazalo da dolazi do semantičkoga usmjeravanja ako su usmjerivač i ciljna riječ međusobno semantički i/ili asocijativno povezani (za pregled vidi Hutchison, 2008 i Lucas, 2009). Slično se utvrdilo i u različitim istraživanjima s dvojezičnim ispitanicima. U tim su istraživanjima usmjerivač i ciljna riječ bili međusobni prijevodi na istim ili različitim pismima (Huynh i Witzel, 2018; Dimitropoulou, Duñabeitia i Carreiras, 2011; Ferré i sur.; 2017; Grainger i Frenck-Mestre, 1998; Lee, Jang, Choi, 2018; Nakayama, Lupker i Itaguchi, 2018), semantički i asocijativno povezani (Basnight-Brown i Altarriba, 2007; Bogunović i Ćoso, 2019; Gulan, 2016; Perea, Duñabeitia i Carreiras, 2008) ili riječi koje dijele značenje i sličnoga su oblika – kognati (Brenders, van Hell i Dijkstra, 2011; Lemhöfer i Dijkstra, 2004; van Hell i Dijkstra, 2002).

U ovom se istraživanju potvrdila činjenica da se međusobno semantički i asocijativno povezane riječi brže prepoznaju u odnosu na one parove usmjerivača i ciljne riječi koji to nisu. To je u skladu s dosadašnjim istraživanjima (Bogunović i Ćoso, 2018; Guasch i sur., 2011; Perea, Duñabeitia i Carreiras, 2008). Teorijsko objašnjenje za tu pojavu može se naći u dvojezičnom interaktivnom aktivacijskom modelu BIA+ (Dijkstra i van Heuven, 2002). Budući da ovaj model polazi od toga da su značenja riječi međusobno povezana čvorovima u zajedničkoj semantičkoj ili leksičkoj mreži, prepoznavanje određene riječi aktivira odgovarajući čvor u mreži. Ta se pak aktivacija dalje automatski širi mrežom pobuđujući bliske, semantički povezane čvorove. Prema tome, ciljna će riječ biti jače aktivirana ukoliko je usmjerivač semantički povezan jer su međusobno bliži nego kada ciljnoj riječi prethodi semantički nepovezana riječ. Budući da je ciljna riječ jače aktivirana ako je semantički povezana s usmjerivačem, brže će se i prepoznati, a vrijeme reakcije će biti kraće.

U jednojezičnom se okruženju pokazalo da dolazi do veće aktivacije dijelova mozga koji se nalaze u čeonom (frontalnom), sljepoočnom (temporalnom) i tjemenom (parijentalnom) režnju ukoliko ciljnim riječima prethode nepovezani usmjerivači u zadatku leksičke odluke u kojem usmjerivači nisu zamaskirani (Ulrich i sur., 2013). Razlog tomu jest semantičko uparivanje (*semantic matching*). Semantičko uparivanje predstavlja proces u

kojem se značenje ciljne riječi uspoređuje sa značenjem usmjerivača (Neely, Keefe i Ross, 1989). Zbog toga ispitanicima treba više vremena da prepoznaju nepovezane ciljne riječi u odnosu na povezane.

U slučaju smjera jezika pokazalo se da se ciljne riječi prepoznaju brže kada im je prethodio usmjerivač na njemačkom jeziku. Kako Huynh i Witzel (2018) u svojem članku tvrde, trenutačna literatura o semantičko-asocijativnom usmjeravanju nije jednoznačna. Za neujednačene se dvojezične osobe smatra da imaju bogatiji semantički prikaz na J1 nego na J2 (Duyck i Brybaert, 2004; Schoonbaert i sur., 2009). To znači da riječ na ispitanikovu J1 aktivira više bliskih konceptualnih čvorova u semantičkoj mreži nego da je ona na J2. Zbog toga se ciljne riječi na J2 prepoznaju brže ako su im prethodili usmjerivači na J1. Kod ujednačenih dvojezičnih osoba polazi se od toga da se aktivira jednak broj bliskih konceptualnih čvorova u semantičkoj mreži (Guasch i sur., 2011). Isto tako valja istaknuti da se mentalni leksikon stalno mijenja ovisno o tome koliko je govornik izložen stranom jeziku (Navracsics, 2000). U kojem će smjeru biti brže usmjeravanje ovisi o vremenu prikaza usmjerivača (Perea, Duñabeitia i Carreiras, 2008), dobi usvajanja jezika (Izura i Ellis, 2002; Silverberg i Samuel, 2004), vrsnosti dvojezičnih govornika (Basnight-Brown i Altarriba, 2007; Meade, Midgley i Holcomb, 2018), uporabi kognata (de Groot i Nas, 1991) ili udjelu konkretnih odnosno apstraktnih riječi (Schoonbaert i sur., 2009 ; van Hell i de Groot, 1998).

U našem se istraživanju pokazalo da dvojezični hrvatsko-njemački ispitanici prepoznaju brže ciljne riječi koje su na hrvatskom ako im je prethodio njemački usmjerivač. Slično se dogodilo i u istraživanju koje su provele Basnight-Brown i Altarriba (2007). U zadatku leksičke odluke s usmjeravanjem ispitanici su bili vrlo vrsni dvojezični govornici španjolskoga i baskijskoga jezika. U njihovu eksperimentu usmjerivač nije bio maskiran tako da su ga ispitanici mogli svjesno percipirati. Pokazalo se da dolazi do usmjeravanja u smjeru baskijski kao J2 - španjolski kao J1. Isto su tako ustanovile da je prepoznavanje bilo brže ukoliko su usmjerivač i ciljna riječ bili prijevodni ekvivalenti u odnosu na semantički povezane usmjerivače i ciljne riječi. Takav se učinak ponovio u drugom eksperimentu s prijevodnim ekvivalentima samo što je usmjerivač u ovom slučaju bio maskiran. Zaključuju da su prijevodni ekvivalenti drugačije pohranjeni u mentalnom leksikonu, nego semantički povezane riječi.

Naši ispitanici također pokazuju iznimno veliku vrsnost u njemačkom jeziku prema samoprocjeni poznavanja njemačkoga jezika. Procijenili su svoje poznavanje njemačkoga

jezika u vještini čitanja, govora i slušnoga razumijevanja iznimno velikim na ljestvici od jedan do pet. Taj je raspon odabran ne bi li bio u skladu s uobičajenim školskim sustavom vrednovanja uspjeha na koji je većina ispitanika naviknula (čitanje: 4.625, govorenje: 4.56, slušanje: 4.69). Valja isto tako istaknuti da su ispitanici često izloženi njemačkom jeziku. To je posljedica toga što su ispitanici u velikom broju nastavnici njemačkoga jezika. Gotovo je polovica ispitanika boravila u zemljama njemačkoga govornog područja ($N = 25$; $M = 8.65$ godina, $SD = 5.83$).

Basnight-Brown i Altarriba (2007) ističu da bi dob usvajanja jezika također mogla biti bitna. Sličnoga su stava Silverberg i Samuel (2004). U svojem su istraživanju uspoređivali rane dvojezične osobe, kasne dvojezične osobe i dvojezične osobe koje su učile J2 u formalnom okruženju. Pokazalo se da kod dvojezičnih osoba koje su rano usvojile svoj J2 dolazi do usmjeravanja u smjeru J2-J1 ako su usmjerivač i ciljna riječ međusobno semantički povezani. Smatraju da do toga dolazi jer je u času kada ciljna riječ bude vidljiva na zaslonu dio njezine konceptualne reprezentacije već aktiviran usmjerivačem. Silverberg su i Samuel tu pojavu uočili jedino kod ranih dvojezičnih osoba koje su J2 usvojili prije 7. godine života. Ispitanici su u našem istraživanju J2 počeli usvajati prosječno u dobi od 8.778 ($SD = 3.489$) što je vrlo blizu navedenom periodu. Silverberg i Samuel ne mogu jasno reći zašto je do takve pojave došlo, no pretpostavljaju da je dječji semantički odnosno konceptualni svijet različit od onoga odraslih osoba.

Silverberg i Samuel (2004) tvrde nadalje da postoje razlike između triju razina modela BIA+ (Dijkstra i van Heuven, 2002). Dijkstra i van Heuven polaze od toga da su kod vrlo vrsnih dvojezičnih govornika ortografska odnosno fonološka razina, leksička razina i konceptualna razina za J1 i J2 međusobno podijeljeni. Međutim, rezultati Silverberga i Samuela ne potvrđuju tu činjenicu u cijelosti. Pokazalo se da su te tri razine samo djelomično međusobno podijeljene. Što će pak biti podijeljeno, ovisi o dobi kada je usvojen J2. Zaključuju da su ispitanici u istraživanju Dijkstre i van Heuvena uglavnom bili vrlo vrsni dvojezični govornici koji su svoj J2 usvojili kasnije u životu. Takvi ispitanici pokazuju konzistentne rezultate koji često ne variraju. No valja uzeti u obzir činjenicu da nisu razmatrani drugi ispitanici koji su pod drugim uvjetima usvojili svoj J2. Zbog toga Silverberg i Samuel zaključuju da ne bi trebalo generalizirati, tvrdeći da dolazi do dijeljenja na svim razinama i da će svi dvojezični govornici polučiti isti rezultat. S druge strane, Perea, Duñabeitia i Carreiras (2008) u svojem su istraživanju otkrili da vrsnost dvojezičnoga

govornika ne utječe na učinak usmjeravanja. On se također pojavljuje kod kasnih vrlo vrsnih dvojezičnih govornika.

Naše je istraživanje također u skladu s nekim postavkama revidiranoga hijerarhijskog modela (RHM) (Kroll i Stewart, 1994). Taj model polazi od toga da se kod neujednačenih i ranih dvojezičnih osoba konceptima pristupa preko J1 u početnim stadijima učenja J2. Kako dvojezični govornik napreduje u svojem poznavanju J2 i stječe veću vrsnost u njemu, može i izravno pristupiti konceptima. Naši su ispitanici, prema tome, mogli izravno pristupiti konceptima. Još nije jasno određena prekretnica kada pojedini dvojezični govornik prestaje biti početnikom i počinje biti vrsnim govornikom. Nasuprot tome, postoje istraživanja koja potvrđuju da i dvojezični govornici koji tek počinju učiti J2 mogu izravno pristupiti konceptima (Duyck i Brysbaert, 2004; Meade, Midgley i Holcomb, 2018). Schoonbaert i sur. (2009) ističu da je model RHM problematičan u tumačenju učinka usmjeravanja ukoliko su usmjerivač i ciljna riječ na različitim jezicima. Ističu kako su Kroll i Stewart (1994) tu pojavu iskoristili kao dokaz jakim vezama između leksikona dvaju jezika.

Dob usvajanja riječi kod dvojezičnih govornika istražili su Izura i Ellis (2002). U svojem su istraživanju pokušali otkriti na koje riječi ispitanici reagiraju brže: na riječi koje se usvajaju rano na J1, ali kasno na J2 ili obrnuto. Polazili su od činjenice da se riječi koje se usvajaju rano u životu brže prepoznaju od onih koje su usvojene kasnije (Morrison i Ellis, 2000). U 3. su i 4. eksperimentu Izura i Ellis (2002) ustanovili da se brže prepoznaju one riječi koje su ranije naučene na J2 bez obzira na to kada su usvojene na J1. Bitno je, dakle, kojim su se redoslijedom riječi usvajale unutar pojedinoga jezika, a ne kako su se usvajali koncepti pojedinih riječi. To je u skladu s postulatima modela RHM kako postoje dva odvojena leksikona, ali zajednički konceptualni prostor. Zbog toga smo preuzeli ispitni materijal u našem istraživanju iz dvaju udžbenika njemačkoga jezika razine A1 prema *Zajedničkom europskom referentnom okviru za jezike (ZEROJ)* (Vijeće Europske unije, 2001). To su *Netzwerk A1* (Dengler i sur., 2017) i *Studio D A1* (Demme, Funk i Kuhn, 2016). Radi se, dakle, o osnovnom vokabularu koji se uči vrlo rano ne samo u nastavi njemačkoga kao stranoga jezika već i u djetinjstvu.

De Wilde, Brysbaert i Eyckmans (2019a) potvrđuju svojim istraživanjem s preko petsto djece školskoga uzrasta rezultate Izure i Ellisa (2002). Utvrdili su značajan učinak dobi kada je pojedina riječ usvojena na J1. Riječi na J1 koje su usvojene u ranijoj dobi jednostavnije će se usvojiti i na J2. Brysbaert i Ellis (2016) tu pojavu tumače pomoću

semantičke hipoteze o učinku dobi usvajanja na J1. Prema toj su hipotezi riječi ili koncepti koji su usvojeni u ranijoj dobi u središtu semantičke mreže zbog čega im dvojezični govornik može lakše pristupiti. Činjenica da se rano usvojene riječi nalaze u središtu semantičke mreže uzrokuje brže učenje riječi na J2.

Jedan od čimbenika koji također može utjecati na brzinu prepoznavanja ciljne riječi jest vrijeme prikaza usmjerivača. Schoonbaert i sur. (2009) ustanovili su u 4. eksperimentu s usmjerivačima na J2 i ciljnim riječima na J1 da je učinak usmjeravanja jači s vremenom prikaza od 250 ms nego od 100 ms. U našem je istraživanju usmjerivač bio prikazan 55 ms. Kako između njega i trenutka kada se pojavila ciljna riječ bio uključen prazan zaslon od 250 ms, došlo je do učinka duhova (*ghosting effect*) (Finkbeiner, 2005) ili ikoničke postojanosti (*iconic persistence*) (Coltheart, 1980). To znači da je usmjerivač prividno duže prikazan nego što u stvarnosti jest. Polazi se od toga da bi to dodatno vrijeme bilo potrebno da dvojezični ispitanici obrade usmjerivač koji bi potom utjecao na prepoznavanje ciljne riječi (Huynh i Witzel, 2018). Prema tome, naši su ispitanici imali dovoljno vremena da procesiraju usmjerivače i na taj način brže reagiraju na ciljne riječi na J1.

U modelu BIA+ (Dijkstra i van Heuven, 2002) važna je razina aktivacije mirovanja riječi (*resting level activation*). Ona predstavlja funkciju kojom se određuje koliko se često pristupa određenoj riječi. Prema tome, razina je aktivacije mirovanja viša kod riječi s većom frekvencijom (Nakayama, Lupker i Itaguchi, 2018). Kako ranije usvojene riječi dvojezični govornik susreće češće, brže će aktivirati čvorove tih riječi u semantičkoj mreži (Brysbaert i Ellis, 2016). Budući da u modelu BIA+ postoji jedinstven leksikon za oba jezika kojima se služi dvojezični govornik, brže će se aktivirati susjedni, semantički bliski čvorovi. Zbog toga će ispitanici brže reagirati na ciljne riječi.

11. EKSPERIMENT 2: ZADATAK LEKSIČKE ODLUKE S USMJERAVANJEM S ISPITNIM MATERIJALOM GENERIRANOGA POMOĆU DSM-A *word2vec*

11.1. Metoda

11.1.1. Ispitanici

U eksperimentu sudjelovalo je sveukupno 44 sudionika (34 ženskih i 10 muških). Ispitanici su bili različite dobi ($M = 36.77$ godina, $SD = 9.41$). Najmlađi su ispitanici bili 21, a najstarija ispitanica 71 godina. Isto kao i u prvom eksperimentu ispitanici odudaraju od uobičajenih ispitanika u ovakvim istraživanjima. Kao i u prvom eksperimentu većina se ispitanika koristi njemačkim u profesionalne svrhe gotovo svaki dan jer su nastavnici i učitelji njemačkoga jezika u javnim i privatnim školama i učilištima ili rade u tvrtkama koje su orijentirane na njemačko govorno područje ili dolaze s njega.

Većina je ispitanika visoke stručne spreme ($N = 37$). Također je većini ispitanika njemački prvi strani jezik ($N = 30$), a ostalima je engleski prvi strani jezik ($N = 14$). Ne postoji značajna razlika u broju godina koji su ispitanici učili njemački jezik ($t(15.297) = .766$, $p > .05$). Ispitanici su njemački počeli učiti s prosječno $M = 8.614$ godina, $SD = 3.301$. Njemački se najčešće učio u školama (54.55%), pomoću televizijskih filmova i serija (22.73%), u stranoj zemlji (13.64%), u obiteljskom okruženju (6.82%) ili u nekom drugom okruženju (2.26%). Prema vlastitim iskazima najviše je njihovu poznavanju njemačkoga jezika doprinijelo čitanje, gledanje televizijskih serija i filmova te samostalno učenje, a najmanje radijske emisije i slušanje glazbe te prijatelji i obitelj. Trenutačno je najviše ispitanika izloženo njemačkom jeziku kroz čitanje, televizijske emisije i internet te kroz radijske emisije, u nešto manjem broju kroz glazbu te samostalno učenje. Najmanje je pak izloženo njemačkom prilikom druženja s prijateljima i obitelji. Da su ispitanici podjednako vrsni u njemačkom jeziku, potvrđuje i činjenica da ne postoji značajna razlika između samoprocjene u vještinama čitanja ($t(16.803) = .642$, $p > .05$), govorenja ($t(19.03) = .537$, $p > .05$) te slušnoga razumijevanja ($t(16.803) = 1.191$, $p > .05$) što se utvrdilo upitnikom koji su ispitanici ispunjavali nakon glavnoga dijela eksperimenta. Isto tako ne postoji značajna razlika između izloženosti njemačkom i engleskom jeziku ($t(43) = 1.166$, $p > .05$) unatoč tomu što su kroz medije gotovo neprestano izloženi engleskom jeziku. Svi se sudionici koriste

redovito njemačkim i hrvatskim jezikom. Svi su ispitanici dobroga vida ili im je ispravljen. Nitko od ispitanika nema jezične ili neurološke poremećaje.

11.1.2. Instrumenti i materijal

Instrumenti kojima je izveden ovaj eksperiment ostali su isti kao u prvom eksperimentu. Promijenili su se, međutim, dijelovi materijala. I dalje postoje dvije liste od 80 parova riječi. Jednu su listu činili parovi usmjerivača i ciljnih riječi u kojima je usmjerivač bio na njemačkom, a ciljna riječ na hrvatskom, a drugu listu činili su parovi u kojima je jezik usmjerivača i ciljne riječi bio obrnut. Polovicu ciljnih riječi svake liste činile su pseudoriječi. Njima su prethodili usmjerivači koji su postojeće riječi na njemačkom ili hrvatskom jeziku. Pseudoriječi oblikovane su pomoću programa *Wuggy* (Keuleers i Brysbaert, 2010).

Isto kao u prvom eksperimentu prvo se utvrdio koncept kojem je trebala odgovarati konkretna imenica (vidi Sliku 16). Potom se odredila riječ koja najviše odgovara tom konceptu u hrvatskom odnosno njemačkom jeziku. Te su riječi onda poslužile kao ciljne riječi. Prema tome, iste su ciljne riječi upotrijebljene u ovom i u prvom eksperimentu. Usmjerivači su se pak određivali u programu *R* (The R Core Team, 2018) pomoću programskoga paketa *LSAfun* te pomoću semantičkoga prostora *dewac_cbow* za njemačke riječi i *hr_cbow* za hrvatske riječi (Günther, Dudschig i Kaup, 2015). Pomoću paketa *LSAfun* može se odrediti semantička udaljenost dane riječi prema svim ostalim riječima u danom semantičkom prostoru. Tako smo za usmjerivač uzeli prvu riječ koja je imala najmanju semantičku udaljenost od ciljne riječi odnosno koncepta osim ako se nisu podudarale ili je pripadala drugoj vrsti riječi. Izbjegavali smo uzeti one riječi koje su prijevodi ciljne riječi. Nepovezane su riječi iste kao u prvom eksperimentu.

Dvojezični se korpus u ovom eksperimentu nije odabrao jer se u prethodnim istraživanjima (Vulić i Moens, 2016) pokazalo da DSM-ovi u takvom korpusu dobivene riječi iz oba jezika međusobno ispremještaju prema sličnosti. Tako je u istraživanju Vulića i Moens s talijansko-engleskim korpusom najbližnja riječ talijanskoj riječi *madre* engleska riječ *mother* koju slijede *padre, moglie, father, sorella...* Zbog toga takav korpus ne bi bio dobar u psiholingvističkim istraživanjima jer bi se u oblikovanju zadatka leksičke odluke s usmjeravanjem međusobno mijenjali jezici usmjerivača i ciljne riječi. Isto se tako iz primjera vidi da je najbližnja riječ talijanskoj riječi *madre* njezin engleski prijevod *mother*.

Semantički je prostor *hr_cbow* (Günther, Dudschig i Kaup, 2015) semantički prostor kontinuiranoga multiskupa riječi (*cbow space*) od 300 dimenzija. Napravljen je na temelju korpusa *OpenSubtitles 2018* (vidi internetsku stranicu www.opensubtitles.org). Taj se korpus sastoji od filmskih i televizijskih podslova na hrvatskom jeziku i ima 707 milijuna pojava. Veličina je korpusa razlog tomu što semantički prostor ima samo 300 dimenzija, a ne 400 kao onaj koji su izradili Baroni, Dinu i Kruszewski (2014). Njihov je korpus, naime, sadržavao 1.5 milijardi pojava. Semantički je prostor generiran algoritmom za CBOW koji je implementiran u model *word2vec* (Mikolov i sur., 2013a; Mikolov i sur., 2013b). Kontekstna je veličina prozora pet riječi. Upotrijebljeno je negativno uzorkovanje od $k = 10$ te poduzorkovanje od $t = 1e - 5$. Semantički prostor *hr_cbow* sadrži sve pojavnice koje se pojavljuju najmanje 50 puta u korpusu *OpenSubtitles 2018*. To čini ukupno 184979 različitih riječi.

Pokazalo se da *R* ima poteškoća s hrvatskim dijakritičkim znakovima na pojedinim grafemima jer nisu dio kôda UTF-8. *R* ih zamjenjuje njihovim inačicama bez dijakritičkih znakova. Kako bi se ti grafemi razlikovali jedni od drugih, zamijenjeni su kombinacijom simbola i grafema (npr. *č* je zamijenjeno sa *_c_*, *ć* sa *_c2_* i sl.).

Semantički prostor *dewac_cbow* (Günther, Dudschig i Kaup, 2015) također je prostor kontinuiranoga multiskupa riječi (*cbow space*). Korpus *deWaC* (Baroni i sur., 2009) poslužio je za izradu semantičkoga prostora *dewac_cbow*. Budući da ovaj korpus ima 1.7 milijardi pojava, semantički prostor *dewac_cbow* ima 400 dimenzija. Taj je korpus nastao na temelju njemačkih internetskih stranica s domenom *.de*. Sastoji se od riječi srednje frekvencije, a polazilo se od korpusa novine *Süddeutsche Zeitung* i popisa riječi osnovnoga vokabulara njemačkoga jezika. Korpus je lematiziran i označene su pojedine vrste riječi (*POS tagged*). Semantički je prostor generiran algoritmom za CBOW koji je implementiran u model *word2vec* (Mikolov i sur., 2013a; Mikolov i sur., 2013b). Kontekstna je veličina prozora pet riječi. Upotrijebljeno je negativno uzorkovanje od $k = 10$ te poduzorkovanje od $t = 1e - 5$. U ovom se semantičkom prostoru nalaze sve one riječi koje se pojavljuju najmanje 50 puta u korpusu. Ukupno postoji 342720 različitih riječi.

Pseudoriječi su iste kao i u prvom eksperimentu. Ne postoji značajna razlika između ciljnih riječi na njemačkom jeziku i usmjerivača na hrvatskom ($t(39) = -1.404$, $p = .168$), no postoji razlika između ciljnih riječi na hrvatskom jeziku i usmjerivača na njemačkom jeziku

($t(39) = -2.167, p = .036$).⁴ Međutim, zbog kriterija kojim smo se vodili pri odabiru ispitnoga materijala, nismo mogli mnogo utjecati na to. Primjere ciljnih riječi i usmjerivača kao i prosječan broj slova po riječi mogu se vidjeti u Tablici 10 i Tablici 11.

Tablica 10: *Primjeri parova usmjerivača i ciljne riječi u eksperimentu s ispitnim materijalom generiranoga pomoću računalnoga programa word2vec. Ciljnoj je riječi uvijek prethodio usmjerivač koji je trebao biti najbliža riječ ciljnoj riječi u semantičkom prostoru hr_cbow i dewac_cbow ili semantički i asocijativno nepovezana riječ. Jedna polovica parova koju su vidjeli ispitanici bila je povezana, dok druga polovica nije. Svaki je ispitanik vidio svaku ciljnu riječ samo jednom.*

Njemačko-hrvatski parovi		
ciljna riječ	usmjerivač (word2vec)	usmjerivač (nepovezana riječ)
Schaf Wolke Sand Ohr	koza sunce tlo grlo	porotnik kopito kos govedo
Hrvatsko-njemački parovi		
ciljna riječ	umjerivač (word2vec)	usmjerivač (nepovezana riječ)
juha jaje jastuk polica	Nudeln Küken Decke Schrank	Kunde Leber Ball Kalb

Tablica 11: *Prosječan broj slova po riječima u eksperimentu s ispitnim materijalom generiranoga pomoću računalnoga programa word2vec. U zagradi su standardne devijacije.*

Njemačko-hrvatski parovi		
ciljna riječ	usmjerivač (word2vec)	usmjerivač (nepovezana riječ)
4.7 (1.018)	5.3 (1.381)	5.0 (.906)
Hrvatsko-njemački parovi		
ciljna riječ	usmjerivač (word2vec)	usmjerivač (nepovezana riječ)
5.275 (1.467)	5.65 (1.889)	5.825 (1.318)

⁴ Ako se uzme u obzir odnos fonema, a ne grafema, onda ne postoji signifikantna razlika između hrvatske ciljne riječi i njemačkoga usmjerivača ($t(39) = .385, p = .702$).

Omjer riječi i pseudoriječi bio je .50. U svakoj je od listi omjer povezanih i nepovezanih parova bio .25. Jedna polovica parova bila je povezana, dok druga nije. Svaki je ispitanik vidio svaku ciljnu riječ samo jedanput. Usmjerivači su se izmjenjivali. Jezik parova usmjerivača i ciljne riječi također se izmjenjivao među ispitanicima. Otprilike je jednak broj ispitanika vidio svaku od dviju lista. To znači da je 21 ispitanik vidio prvo njemačko-hrvatske parove, a potom hrvatsko-njemačke parove, dok je 23 ispitanika vidjelo prvo hrvatsko-njemačke parove, a potom njemačko-hrvatske parove. Cjelovit popis ciljnih riječi s povezanim i nepovezanim usmjerivačima te pseudoriječima može se naći u prilogima koji prate ovaj rad.

11.1.3. Postupak

Postupak je bio isti kao u prvom eksperimentu. To znači da se zbog otežane mogućnosti okupljanja i nemogućnosti ulaska u škole, fakultete i slične ustanove ispitanike trebalo naći na drugačiji način. U prikupljanju podataka istraživaču su pomogli prijatelji. Podaci su se centralno skupljali pomoću alata *Just Another Tool for Online Studies* – JATOS (Lange, Kühn i Filevich, 2015) na internetskom serveru *Amazon Web Servicesa (AWS)*. Taj alat dodjeljuje zasebnu internetsku adresu svakom ispitaniku pomoću koje pristupa istraživanju koje se potom skida i lokalno čuva na računalu. Alat ne čuva ispitanikovu IP-adresu, pa se tako čuva njegova anonimnost. Kad ispitanik završi s eksperimentom, podaci se pohranjuju na serveru odakle se mogu skinuti u *txt*-formatu. Pomoću programa *OpenSesame* (Mathôt, Schreij i Theeuwes, 2012) podaci se pretvaraju u *csv*-format nakon čega ih se može statistički obraditi.

Kao u prvom eksperimentu i u ovom smo se koristili zadatkom leksičke odluke s usmjeravanjem. To znači da je svakoj ciljnoj riječi prethodio usmjerivač. Usmjerivač je mogao biti povezan s ciljnom riječi ili nepovezan. Povezanost se pak utvrdila semantičkom udaljenosti između usmjerivača i ciljne riječi. Što je semantička udaljenost manja, usmjerivač su i ciljna riječ sličniji. Suprotno tome, što je semantička udaljenost veća, usmjerivač su i ciljna riječ različitiji. Usmjerivač i ciljna riječ nisu bili na istom jeziku.

Upute ispitanicima tijekom istraživanja bile su napisane hrvatskim jezikom na zaslonu računala. Korištena su crna slova fonta Liberation Sans veličine 30 na bijeloj podlozi. Ispitanici su upoznati s time da se istraživanje sastoji od dva dijela. Prvi je dio bio na

hrvatskom, a drugi na njemačkom odnosno obratno. Ispitanicima su dani primjeri za postojeće i nepostojeće riječi na hrvatskom odnosno njemačkom jeziku.

Na Slici 17 može se vidjeti kako je tekao eksperiment. Na početku bi se pojavila točka fiksacije kojom se odvajao svaki par usmjerivača i ciljne riječi od sljedećega. Bila je vidljiva 750 ms. Nakon nje pojavio bi se usmjerivač u trajanju od 55 ms. Uslijedio je međupodražajni interval od 250 ms nakon kojega je slijedila ciljna riječ. Na zaslonu se zadržala najviše dvije sekunde. Kako nije bilo maskiranja, ispitanici su bili upozoreni na to da će prije riječi na koju trebaju reagirati biti kratko vidljiva riječ na drugom jeziku.

Ispitanici su trebali pritisnuti tipku "j" ako je riječ na zaslonu postojeća riječ na hrvatskom ili njemačkom jeziku. U suprotnom su slučaju, ako riječ na zaslonu nije bila postojeća riječ na jednom od dvaju jezika, trebali pritisnuti tipku "f". Prije nego što je eksperiment započeo, ispitanici su bili zamoljeni da postave svoj desni kažiprst na tipku "j", a svoj lijevi kažiprst na tipku "f". Čim bi ispitanik pritisnuo jednu od tih dviju tipki, eksperiment bi krenuo i pojavio bi se prvi par ispitnih riječi. Ispitanici su trebali reagirati što točnije i brže pritiskom na odgovarajuću tipku. Ako su ispitanici pogrešno odgovorili na podražaj, na zaslonu bi vidjeli riječ "netočno" napisanu crvenim slovima. Ondje bi se zadržala 750 ms.

Svaki je blok eksperimenta krenuo s 20 parova usmjerivača i ciljnih riječi za vježbu. To znači da su ispitanici imali 20 parova za vježbu ispred hrvatskoga dijela (njemački usmjerivač i hrvatska ciljna riječ) i isto toliko parova ispred njemačkoga dijela (hrvatski usmjerivač i njemačka ciljna riječ). Ispred svakoga su dijela ispitanici bili upoznati s načinom kako će se izvesti eksperiment. Eksperiment je trajao oko 15 minuta.

Kada su ispitanici bili gotovi sa zadatkom leksičke odluke s usmjeravanjem, podaci bi se pohranili na serveru, a ispitanike bi se istodobno preusmjerilo na drugu internetsku stranicu gdje su ispunjavali kratak upitnik o svojoj jezičnoj vrsnosti i biografiji.

U prilogima koji prate rad može se naći detaljniji pregled sličica s objašnjenjima za ispitanike te upitnik koji su ispunjavali.

11.2. Rezultati

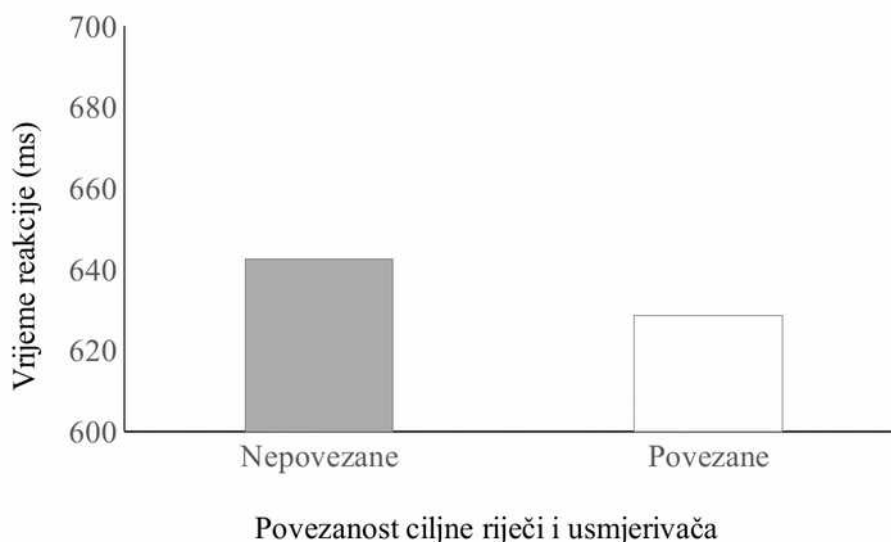
U drugom eksperimentu zabilježeno je ukupno 3.5 % netočnih odgovora koji su isključeni iz daljnjih analiza. Iz analiza su također isključeni odgovori koji su bili kraći od 200 ms (0.06 %) i duži od 1650 ms (0.42 %). Kao i u prvom eksperimentu, prije glavne analize

izračunata su prosječna vremena reakcije za svih 44 sudionika u četiri eksperimentalna uvjeta. U Tablici 12 prikazani su rezultati deskriptivne statistike za vrijeme reakcije s obzirom na povezanost usmjerivača i ciljne riječi te s obzirom na jezik na kojem je zadana ciljna riječ.

Tablica 12: Mjere deskriptivne statistike u drugom eksperimentu za vrijeme reakcije s obzirom na povezanost ciljne riječi i usmjerivača te jezik na kojem je zadana ciljna riječ ($N = 44$).

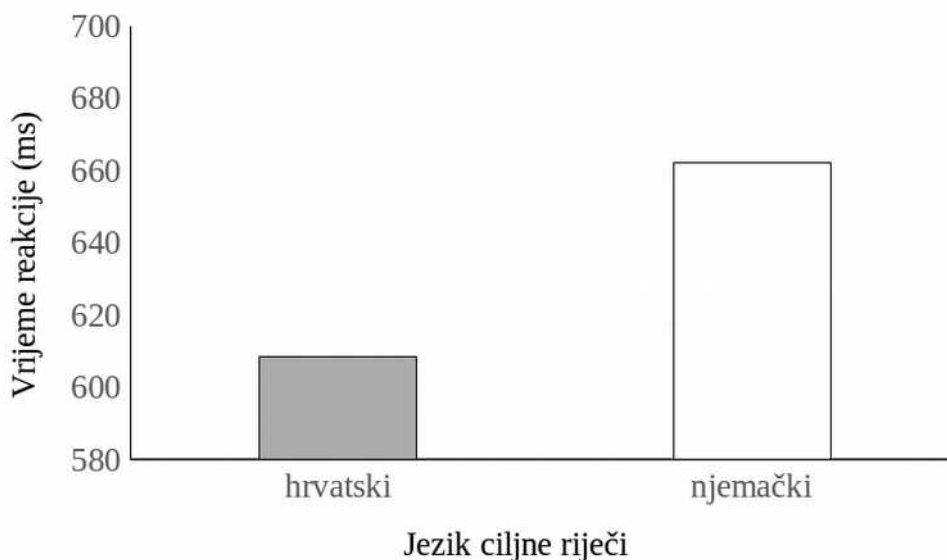
Povezanost riječi	Jezik ciljne riječi	Aritmetička sredina (M)	Standardna devijacija (SD)
nepovezane	hrvatski	614.69	101.27
	njemački	670.30	107.73
povezane	hrvatski	602.59	91.46
	njemački	654.41	99.60

Ponovno je provedena 2×2 dvosmjerna analiza varijance s ponovljenim mjerenjima. Kao i u prvom eksperimentu pronađeni su značajni glavni efekti, dok interakcija između faktora (povezanost \times jezik ciljne riječi) ponovno nije bila značajna. Pronađen je umjereno veliki efekt povezanosti ciljne riječi i usmjerivača: $F(1, 43) = 5.17$; $p = .028$; $\eta^2 = .107$. Sudionici istraživanja su imali brže vrijeme reakcije kada su ciljna riječ i usmjerivač bili povezani pomoću računalnoga programa ($M = 628.50$) nego kada su bili nepovezani ($M = 642.50$). Slika 20 prikazuje glavni efekt povezanosti riječi u drugom eksperimentu.



Slika 20: Prosječne vrijednosti vremena reakcije u drugom eksperimentu s obzirom na povezanost ciljne riječi i usmjerivača.

Pronađen je veliki glavni efekt jezika na kojem je bila zadana ciljna riječ: $F(1, 43) = 31.80$; $p < .001$; $\eta^2 = .425$. Kao i u prvom eksperimentu sudionici su imali brže vrijeme reakcije kada je ciljna riječ bila na hrvatskom, a usmjerivač na njemačkom jeziku ($M = 608.64$) nego u obrnutoj situaciji ($M = 662.36$). Slika 21 prikazuje glavni efekt jezika ciljne riječi u drugom eksperimentu.



Slika 21: Prosječne vrijednosti vremena reakcije u drugom eksperimentu s obzirom na jezik na kojem je zadana ciljna riječ.

11.3. Rasprava

U drugom smo eksperimentu, slično kao u prvome, htjeli ustanoviti postoji li semantičko usmjeravanje između usmjerivača i ciljne riječi koji su na različitim jezicima, hrvatski jezik kao J1 i njemački jezik kao J2. Isto se tako pokušalo odrediti u kojem se smjeru odvija usmjeravanje.

Rezultati su ovoga eksperimenta potvrdili one prvoga. Pokazalo se da dolazi do bržega prepoznavanja ciljne riječi ako joj je prethodio usmjerivač koji je asocijativno povezan. Isto su se tako ponovili rezultati prvoga eksperimenta što se tiče smjera usmjeravanja. Došlo je do bržega prepoznavanja ciljne riječi ukoliko je ona na J1, a usmjerivač na J2.

Prema našim saznanjima do sada nije bilo istraživanja koja su na ovakav način izveli zadatak leksičke odluke s umjeravanjem u dvojezičnom okruženju. Provedeno je niz istraživanja u jednojezičnom okruženju, no glavnina je samo nastojala otkriti potvrđuju li različiti distribucijski semantički modeli (DSM) rezultate dobivene uobičajenim zadacima

leksičke odluke s usmjeravanjem u kojima je istraživač samostalno određivao parove usmjerivača i ciljnih riječi. Pokušalo se odrediti potvrđuje li sličnost između usmjerivača i ciljne riječi njezino brže prepoznavanje (Ettinger i Linzen, 2016; Jones, Kintsch i Mewhort, 2006; Mandera, Keuleers i Brysbaert, 2017; Pereira i sur., 2016; Samir, Beekhuizen i Stevenson, 2020). Međutim, do sada su vrlo rijetko upotrijebljeni DSM-ovi kako bi se generirao ispitni materijal. Stoga je naš rad jedan od prvih u kojem se koristi računalnim modelom *word2vec* kako bi se generirao ispitni materijal u dvojezičnom okruženju. Kako su se prema našim saznanjima do sada računalni modeli uglavnom koristili samo u jednojezičnim istraživanjima, nema radova na koje bismo se mogli referirati. U nastavku želimo prikazati kako se u jednojezičnim istraživanjima odrazilo korištenje računalnih modela prilikom određivanja ispitnoga materijala. Nadalini i sur. (2018) izveli su eksperiment sa zadatkom kategorizacije pri čemu je usmjerivač bio skriven maskom koja mu je prethodila i koja ga je slijedila te eksperiment u kojem usmjerivač nije bio skriven maskama. Vrijeme prikazivanja usmjerivača razlikovalo se među ispitanicima. Trebali su odrediti pripada li ciljna riječ u prvom dijelu eksperimenta kategoriji životinja, a u drugom dijelu kategoriji alata. Sličnost se između usmjerivača i ciljne riječi odredila pomoću statističke mjere točkaste procjene uzajamne informacije (*pointwise mutual information - PMI*) te pomoću kosinusa kuta između dvaju vektora riječi. Kosinus se odredio na temelju korpusa koji je nastao treniranjem modela *word2vec* (Mikolov i sur., 2013a; Mikolov i sur., 2013b). Sličnosti određene pomoću PMI-a i *word2vec*-a nastale su na temelju istoga talijanskog korpusa *ItWaC* koji je bio lematiziran i gramatički tagiran (Nadalini i sur., 2018). Pokazalo se da PMI i *word2vec* imaju značajan učinak na usmjeravanje, no rezultati ukazuju na to da je učinak uzajamne informacije PMI-a jači.

Semantičko usmjeravanje zadatkom leksičke odluke s usmjeravanjem na temelju njemačkoga korpusa istražili su Günther, Dudschig i Kaup (2016a; 2016b). Jedinu su koji su unaprijed odredili ispitni materijal pomoću DSM-a. Uočili su, naime, da se asocijativna sličnost određivala uvijek naknadno što je dovelo do toga da se može manipulirati sličnošću između usmjerivača i ciljnih riječi. To nas je potaknulo da se i mi u ovom eksperimentu koristimo ispitnim materijalom koji smo unaprijed odredili pomoću modela *word2vec*. Günther, Dudschig i Kaup (2016a) u svojem su istraživanju nastojali parove ispitnih riječi ravnomjerno rasporediti preko cijeloga spektra sličnosti te izbjegavati grupiranje u dvjema krajnostima kao što se to dogodilo u istraživanju koje su proveli Hutchison i sur. (2008).

Međutim, naši su povezani parovi usmjerivača i ciljnih riječi upravo bili što udaljeniji od nepovezanih parova ne bismo li na taj način mogli jasnije ustanoviti učinak usmjeravanja.

Günther, Dudschig i Kaup (2016a) nadalje iznose da se sličnost na temelju kosinusa kuta između usmjerivača i ciljne riječi uvijek određivala *post hoc*. Ta je činjenica onemogućavala da se kontroliraju korelacije između drugih varijabli u eksperimentu. Stoga se Günther, Dudschig i Kaup (2016a) zalažu za to da se kosinus kuta kao mjera sličnosti uvede kao dodatna neovisna varijabla pokraj duljine riječi, frekvencije, veličine susjedstva i sl.

U svojem su prvom istraživanju Günther, Dudschig i Kaup (2016a) upotrijebili distribucijski semantički model LSA (Landauer i Dumais, 1997). Opredijelili su se za taj model jer je jedan od najzastupljenijih i najcitiranijih. Kako bi provjerili postoji li učinak koji uzrokuje LSA, odlučili su se da će to ispitati pomoću zadatka leksičke odluke s usmjeravanjem. Nadalje su se odlučili za to da će se koristiti relativno dugim vremenskim razmakom između prikaza usmjerivača i ciljne riječi od 1000 ms (500 ms je bio vidljiv usmjerivač na zaslonu nakon čega je slijedilo dodatnih 500 ms tijekom kojih je zaslon bio prazan). Polazili su od toga da će u tom slučaju učinak usmjeravanja biti značajniji. U našem je eksperimentu vrijeme od trenutka pojavljivanja usmjerivača od trenutka pojavljivanja ciljne riječi iznosilo 305 ms. Takvo je relativno kratko vrijeme u usporedbi s vremenom kod Günthera, Dudschig i Kaup (2016a) također polučilo znakovit učinak usmjeravanja.

Ključno je u istraživanju Günthera, Dudschig i Kaup (2016a) način odabira ispitnoga materijala. Taj smo stav preuzeli i mi. Stoga smo se odlučili da unaprijed odaberemo ispitni materijal za naše istraživanje. Nakon što su iz njemačkoga korpusa uklonili sve zaustavne riječi i sve riječi koje se ne pojavljuju najmanje u dva dokumenta, proveli su uobičajeni postupak pri oblikovanju semantičkoga prostora u LSA-u (za detalje vidi poglavlje o DSM-ovima). Korpus se sastojao od približno pet milijuna riječi. Potom su pristupili odabiru ispitnoga materijala. Definirali su četiri kategorije sličnosti ovisno o kosinusu kuta između usmjerivača i ciljne riječi. Kategorije su činile vrijednosti kosinusa kuta između (i) .00 i .10, (ii) .10 i .25, (iii) .25 i 40 te (iv) .40 i .60. Budući da vrlo sličnih riječi ima iznimno malo, a riječi koje nisu slične ima vrlo mnogo, odlučili su se za takvu podjelu kategorija. Drugačije postavljeno, riječi koje nisu slične ima mnogo, stoga je raspon vrijednosti kosinusa uži. S druge strane, raspored je vrijednosti kosinusa širi jer ima malo riječi koje su vrlo slične. Sličnost se određivala pomoću paketa *LSAfun* (Günther, Dudschig i Kaup, 2015) u programskom okruženju *R* (R Core Team, 2018). Günther, Dudschig i Kaup (2016a) odredili

su niz usmjerivača kojima su pomoću programskoga paketa *LSAfun* odredili ciljne riječi. Ukoliko je usmjerivač bio dodijeljen prvoj kategoriji sličnosti, odredili su ciljnu riječ tako da su uzeli prvu riječ koja se nalazila unutar vrijednosti kosinusa između .00 i .10. Takva je ciljna riječ trebala zadovoljavati nekoliko dodatnih uvjeta: morala je biti konkretna imenica, morala je imati određen broj slogova i slova te određenu logaritamsku frekvenciju. Ako je dotična riječ zadovoljila navedene uvjete, uzeta je kao ciljna riječ. Takav su način odabira ispitnih riječi istraživači nazvali pseudonasumični pristup. U drugom su eksperimentu zamijenili usmjerivače i ciljne riječi. Pojednostavljeno rečeno, riječ koja je u prvom eksperimentu bila usmjerivač, u drugom je postala ciljna riječ i obrnuto, riječ koja je u prvom eksperimentu bila ciljna riječ u drugom je postala usmjerivač. S tim su eksperimentom htjeli provjeriti postoji li učinak unazadnoga usmjeravanja (*backward priming effect*). Taj je učinak prvi uočio Koriat (1981). Pojava je to u kojoj, primjerice, usmjerivač *roda* može pospješiti prepoznavanje ciljne riječi *beba* (usmjeravanje unaprijed), no to se neće uvijek dogoditi i u obrnutom smjeru gdje je usmjerivač *beba*, a ciljna riječ *roda* (usmjeravanje unazad) (za pregled vidi Hutchison, 2003). U oba se eksperimenta razlika u kosinusu kuta, utvrđena LSA-om, pokazala znakovitom varijablom. Veća kosinusna sličnost između usmjerivača i ciljne riječi predviđa kraće vrijeme reakcije. Iako smo uvidjeli važnost odabira ispitnoga materijala unaprijed, na što su nas potaknuli radovi Günthera, Dudschig i Kaup (2016a; 2016b), nismo se mogli voditi istim kriterijem kao oni jer je naše istraživanje bilo dvojezično i nismo se fokusirali na učinak unazadnoga usmjeravanja.

Günther, Dudschig i Kaup (2016b) potvrdili su svoje zaključke dobivene na temelju modela LSA i proširili ih na model HAL. U svojem drugom istraživanju nisu samo uključili dodatni DSM već su upotrijebili i korpus koji je znatno veći nego u prvom istraživanju. Korpus *sDeWaC* (Faaß i Eckart, 2013), kojime su se koristili u drugom istraživanju, sastoji se od otprilike 880 milijuna pojavnica. Veći korpus i samim time veća količina podataka poboljšava i preciznije određuje sličnosti pomoću kosina kuta između dviju riječi (Bullinaria i Levy, 2007; Bullinaria i Levy, 2012). Stoga Günther, Dudschig i Kaup (2016b) pretpostavljaju kako je moguće da su kosinusne sličnosti utvrđene u eksperimentima koji su proveli Günther, Dudschig i Kaup (2016a) ili precijenjene ili podcijenjene.

U svojem su prvom eksperimentu Günther, Dudschig i Kaup (2016b) konstruirali semantički prostor pomoću modela LSA. No ovaj su puta također povećali broj kategorija sličnosti na deset (prvu su kategoriju sličnosti tako činile vrijednosti između .00 i .09, drugu

kategoriju sličnosti vrijednosti između .10 i .19 itd.). Na taj su način postigli ujednačeniju distribuciju sličnosti između usmjerivača i ciljnih riječi. Svakoju su ciljnoj riječi potom odredili usmjerivač.

Slično su učinili i u svojem drugom eksperimentu. S korpusom *sDeWaC* (Faaß i Eckart, 2013) konstruirali su semantički prostor pomoću modela HAL i odredili su usmjerivače zadanim ciljnim riječima koje su bile jednake onima u prvom eksperimentu. Pokazalo se da sličnosti generirane kosinusom kuta modela LSA ne polučuju učinak usmjeravanja, dok se u drugom eksperimentu potvrdilo da sličnosti određene modelom HAL mogu polučiti takav učinak. Günther, Dudschig i Kaup (2016b) zaključuju da model HAL bolje predviđa učinak usmjeravanja od modela LSA. Razlog tomu vide u činjenici što su vektori u ta dva modela drugačije kodirani. LSA kodira općenite informacije o tekstu u svojim vektorima jer ima cijeli dokument kao kontekst, dok HAL obuhvaća neposredan kontekst (ograničen broj riječi koji prethodi središnjoj riječi i koji ju slijedi).

Mandera, Keuleers i Brysbaert (2017) u svojem istraživanju usporedili modele prebrojavanja (npr. LSA i HAL) s predviđačkim modelima (npr. *word2vec*) i došli su do zaključka da potonji uglavnom nadmašuju prve. Do toga su zaključka došli analizirajući vrijeme reakcije na velikom skupu podataka semantičkoga usmjeravanja (Hutchison i sur., 2008). Međutim, i u tom su se istraživanju koristili samo jednojezičnim podacima. Prema tome, ne postoji nijedno prijašnje istraživanje s kojim bismo mogli usporediti rezultate dobivene našim istraživanjem. Međutim, možemo ustvrditi da rezultati dobiveni modelom *word2vec*, odražavaju rasuđivanje kod ljudi. Pokazalo se, naime, da materijal generiran na tradicionalan način odabirom istraživača na temelju diskretnih slobodnih asocijacija i materijal generiran pomoću predviđačkoga modela *word2vec* polučuje jednake rezultate u pogledu učinka povezanosti usmjerivača i ciljne riječi i smjera usmjeravanja. Prema tome, kosinus sličnosti među riječima pokazao se u ovom istraživanju kao mjerodavan u određivanju učinka usmjeravanja. Relevantni su parovi u našem istraživanju imali veliku sličnost između usmjerivača i ciljne riječi (za hrvatsko-njemačke parove $M = .444$, $SD = .115$, a za njemačko-hrvatske parove $M = .565$, $SD = .103$) za razliku od nepovezanih parova riječi (za hrvatsko-njemačke parove $M = .048$, $SD = .107$, a za njemačko-hrvatske parove $M = .104$, $SD = .121$). Zbog toga se *word2vec* može smatrati dobrim alatom kojim se može predstaviti značenje riječi, a kosinus kuta između dviju riječi dobrom mjerom za sličnost među riječima. Drugim riječima, budući da veći kosinus kuta između dviju riječi pospješuje prepoznavanje

riječi, tako dobiven učinak može se smatrati učinkom usmjeravanja. Do sličnoga su zaključka došli Günther, Dudschig i Kaup (2016a; 2016b) u svojim istraživanjima. Međutim, potrebna su daljnja istraživanja kako bi se ta činjenica potvrdila s drugim ispitanicima, drugačijim ispitnim materijalom i kombinacijom jezika.

12. USPOREDBA PRVOGA I DRUGOGA EKSPERIMENTA

12.1. Rezultati i rasprava

Kako bismo provjerili istraživačku hipotezu rezonira li računalni model *word2vec* rasuđivanje kod ljudi, najprije su izračunata ukupna prosječna vremena reakcije za 54 sudionika prvog eksperimenta te ukupna prosječna vremena reakcije za 44 sudionika drugoga eksperimenta. Sudionici u eksperimentu u kojem su riječi bile generirane pomoću računalnoga programa *word2vec* imali su nešto brže prosječno vrijeme reakcije ($M = 635.50$; $SD = 90.94$) nego sudionici u eksperimentu s asocijativno povezanim riječima ($M = 664.27$; $SD = 84.43$). Kako bismo provjerili statističku značajnost ove razlike, proveden je nezavisni t-test koji je pokazao da ova razlika nije statistički značajna: $t(96) = 1.62$, $p = .108$. Stoga možemo zaključiti da se sudionici prvoga i drugoga eksperimenta nisu međusobno razlikovali s obzirom na svoje prosječno vrijeme reakcije.

Ustanovili smo da ne postoji statistička razlika između prvoga i drugoga eksperimenta. To je u skladu s idejom koju zastupaju Günther, Rinaldi i Marelli (2019). Smatraju da DSM mora biti u stanju rješavati one zadatke koje obavljaju ljudi i da ih ne smije premašiti. Ponekada DSM-ovi imaju bolje rezultate nego što to ih imaju ljudi (Mandera, Keuleers i Brysbaert, 2017).

Prepoznali su dvije važne odrednice koje mora ispunjavati svaki DSM kako bi valjano mogao odraziti semantičku memoriju, a to su, prvo, da mora uspješno predvidjeti ljudsko ponašanje u nizu različitih zadataka u kojima je semantička memorija bitna, pa čak i u onima za koje nije bio izvorno namijenjen, te, drugo, da pokazuje rezultate koji su slični onima kod ljudi. Pokazalo se da modeli poput *word2vec*-a nadmašuju tradicionalne modele prebrojavanja u nizu zadataka kao što su zadatak ocjenjivanja sličnosti, zadatak kategorizacije riječi, semantičko usmjeravanje i zadatak procjene jesu li dvije riječi sinonimi ili ne (Baroni, Dinu i Kruszewski, 2014; Mandera, Keuleers i Brysbaert, 2017). Osim toga Mandera, Keuleers i Brysbaert (2017) tvrde da su predviđački modeli poput *word2vec*-a prihvatljiviji jer se koriste modelom asocijativnoga učenja koji su predložili Rescola i Wagner (1972) te su zbog toga psihološki mnogo utemeljeniji nego modeli prebrojavanja poput HAL-a ili LSA-a kojima se često služilo u istraživanjima (Hutchison i sur., 2013). U svojem su istraživanju uočili da se različiti DSM-ovi ne razlikuju mnogo jedni od drugih u zadacima leksičke odluke s usmjeravanjem i imenovanja. Nasuprot tome, Baroni, Dinu i Kruszewski (2014) u

poredbenom su istraživanju pokazali kako su predviđački modeli znatno bolji od modela prebrojavanja.

Kako su se predviđački modeli pokazali boljima od modela prebrojavanja te kako su prvi psihološki utemeljeniji, odlučili smo se za to da se njima koristimo pri izradi ispitnoga materijala. Mnogi istraživači, međutim, smatraju kako na uspješnost pojedinoga modela utječu i čimbenici poput odabira korpusa, veličine prozora i sl. (Günther, Rinaldi i Marelli, 2019; Kumar, 2020; Levy, Goldberg i Dagan, 2015; Mandera, Keuleers i Brysbaert, 2017). Iako se ističe da se s većim korpusom, povećava učinak DSM-a (Recchia i Jones, 2009), to nije jednako važno u svim zadacima. Veličina je korpusa tako bitna u testovima poznavanja vokabulara, no nije bitna u zadacima koji ispituju semantičku memoriju poput semantičkoga usmjeravanja ili imenovanja. U tim su se zadacima pokazali boljima korpusi koji su nastali na temelju podslova serija ili čak objava na različitim društvenim mrežama (Brysbaert, Keuleers i New, 2011; Heraĝdelen i Marelli, 2017; Mandera i sur., 2015). Takvi korpusi često bolje odgovaraju jezičnom iskustvu govornika jer jezik filmova, serija i društvenih mreža više odražava svakodnevni jezik nego jezik kojim su pisane knjige. Zbog toga smo se odlučili za korpus koji je nastao na temelju hrvatskih podslova serija i filmova.

Asr, Willits i Jones (2016) usporedili su model prebrojavanja HAL i predviđački model *word2vec*. Ispitali su kako se usvajaju semantičke kategorije na temelju korpusa koji se sastojao od govora koji je bio usmjeren djeci. Njihovi su rezultati pokazali da u takvom korpusu bolje rezultate ima HAL nego *word2vec*. Pokazalo se da DSM-ovi ne pokazuju deficite samo u korpusu usmjerenom djeci već i u tome da ne mogu naučiti nove koncepte u situacijama u kojima su samo jednom izloženi određenom podražaju (Landau, Smith i Jones, 1988). Günther, Rinaldi i Marelli (2019) ističu da postoje računalni modeli koji su u stanju naučiti pojave kojima su izloženi samo jedanput (Lake, Salakhutdinov i Tenenbaum, 2015), no takvi se modeli ne mogu ubrajati među DSM-ove jer počivaju na drugačijim idejama. Tvrdi da se u ovome trenutku ne može jasno reći hoće li se ta dva modela uspjeti spojiti u jedan opći model.

Problem je utemeljenja DSM-a već ranije istaknut (Barsalou, 1999; Glenberg i Robertson, 2000; Barsalou, 2008). Navodi se kako tijelo, mozak i okolina interagiraju i na taj način stvaraju znanje i utječu na ponašanje. To čini jedan od glavnih prigovora DSM-u. Matheson i Barsalou (2018) upozoravaju na to da ne postoji teorija utemeljenja. Ističu da je među postojeće četiri hipoteze o utemeljenju - hipoteza zamjene (*replacement hypotheses*),

hipoteza ustrojstva (*constitution hypotheses*), hipoteza utjecaja (*influence hypotheses*) i hipoteza konceptualizacije (*conceptualization hypotheses*) - najutjecajnija potonja hipoteza konceptualizacije. Smatraju da su simulacije i reaktivacije osjetilnoga i motoričkoga sustava osnove kognicije. Međutim, nepostojanje teorije znači da je nejasno pospješuju li obavijesti koje su specifične za određene modalnosti poput haptičkih, olfaktivnih i dr. tijekom izvršavanja zadataka ispitanikovo ponašanje (Pulvermüller, 2005) ili dovode do inhibicije (Kumar, 2020). Taj se problem pokušao riješiti tako da su se DSM-ovi kombinirali s informacijama koje se dobivaju iz različitih obilježja riječi (*feature information*) (Andrews, Vigliocco i Vinson, 2009; Johns i Jones, 2015). Andrews, Vigliocco i Vinson (2009) kombinirali su semantička obilježja koja se temelje na iskustvenom doživljaju (*experiential data*) poput obilježja <laje>, <ima rep>, <ima raščupanu dlaku>, <lovi mačke> i dr. za riječ *pas* s distribucijskim obilježjima poput onoga da se riječi *pas* i *mačka* često pojavljuju u istom okruženju. Zaključuju kako se znanje o svijetu može nadograditi znanjem koje je pojedinac stekao izloženošću jeziku. Drugim riječima, moguće je da pojedinac nije iskustveno doživio neku stvarnu pojavu ili predmet, međutim, posredno pomoću jezika može zaključivati o obilježjima dotične pojave ili predmeta. De Deyne, Perfors i Navarro (2016) također zastupaju mišljenje da bi podatke koji su dobiveni isključivo iz jezičnoga korpusa (eksterni jezik - *external language* ili *E-language*) trebalo povezati s izvanjezičnim podacima poput slika i sl. U tom bi slučaju rezultati računalnih modela (*E-language model*) bili mnogo sličniji modelima internoga jezika (*I-language model*) koji jezično znanje vide unutar govornikova mozga. Kiela je sa svojim suradnicima izradio multimodalne DSM-ove koji su u uobičajene modele uključili vizualne, zvučne i olfaktivne podatke (Kiela i sur., 2014; Kiela, 2017; Kiela, Bulat i Clark, 2015; Kiela i Clark, 2015).

Nadalje se kritiziralo to što se teško moglo objasniti utemeljenje apstraktnih koncepata kao što je *ljubav*, *razum*, *sloboda* i dr. Lakoff i Johnson (1999) pokušali su utemeljenje apstraktnih koncepata objasniti metaforama. Isto su tako Vigliocco i sur. (2013) pokušali objasniti da su apstraktni koncepti utemeljeni u našim unutarnjim emocionalnim stanjima. Međutim, može se zaključiti da riječi nisu samo utemeljene u osjetilnoj percepciji i motoričkim kretnjama već su i simbolički međusobno povezane što omogućuje lakšu obradu. Takvu je ideju zastupao Louwerse sa svojom hipotezom simboličke međuovisnosti (*symbol interdependency hypothesis*) (Louwerse, 2011; Louwerse i Zwann, 2009). Mnoga istraživanja koja su se koristila DSM-ovima potvrdila su postavke ove hipoteze (Louwerse, 2011;

Hutchison i Louwarse, 2013; Louwarse i Connell, 2011). Budući da se ljudi koriste jezikom kako bi opisivali svijet, jezik ne postoji neovisno o svijetu. Strukture semantičkih veza unutar jezika zrcale strukture svijeta koji nas okružuje. Naposljetku se može istaknuti da, iako takva veza između svijeta i jezika postoji, ona nije savršena (Günther, Rinaldi i Marelli, 2019).

13. PEDAGOŠKE IMPLIKACIJE

U nastavku će se rada iznijeti nekoliko čimbenika koji se često uzimaju u obzir u dvojezičnim istraživanjima, posebice kada se koristi zadatkom leksičke odluke s usmjeravanjem. Vidjet ćemo kako se oni odražavaju na stjecanje vrsnosti. Prema podacima koji su ispitanici dali nakon što su ispunili upitnik, možemo vidjeti da su neki od tih čimbenika vjerojatno utjecali i na njihovo savladavanje njemačkoga jezika do vrlo visoke razine.

Cilj je odgojno-obrazovnoga nastavnog procesa osposobiti govornike da ovladaju stranim jezikom razine približno jednake J1. Ustanovili smo da je to slučaj kod naših ispitanika. To su postigli formalnim obrazovanjem u školi, boravkom u zemljama njemačkoga govornog područja te čestom izloženošću i korištenju njemačkoga jezika. U ovom ćemo poglavlju analizirati kako kod dvojezičnih govornika postići razinu sličnu J1, odnosno kako povećati vrsnost govornika. Do tih se spoznaja došlo psiholingvističkim eksperimentima, poglavito zadatkom leksičke odluke s usmjeravanjem.

13.1. Metoda poučavanja stranoga jezika

Altarriba i Knickerbocker (2011) ističu kako se vrijednost međujezičnoga učinka usmjeravanja može koristiti ne bi li se na taj način procijenila koliko su dobro govornici J2 usvojili taj jezik. Navode kako jače veze između J1 i J2 uzrokuju veći učinak usmjeravanja. Našim smo istraživanjem upravo došli do tog rezultata. Potvrdili smo kako su naši ispitanici vrsni govornici J2.

U svojem su istraživanju Altarriba i Knickerbocker (2011) uspoređivali tri različite metode kojima se može poučavati vokabular stranoga jezika. To su bile metoda poučavanja prijevodom prema kojoj se stranoj riječi pridjenula riječ na J1, metoda crno-bijelih slikovnih kartica kojom se nove ili nepoznate riječi na J2 povezuju s crno-bijelim slikovnim karticama koje prikazuju određeni predmet te metoda slikovnih kartica u boji koja je ista kao prethodna, samo što su slike koje prikazuju određene predmete u boji. Smatraju da tri čimbenika utječu na lakoću usvajanja novoga stranog jezika. To su pozornost, slikovni izvori koji utječu na poboljšano pamćenje informacija i povezivanje novostečenih informacija s postojećim informacijama. Altarriba i Knickerbocker polaze od toga da uparivanje novoga vokabulara sa slikama potiče brže prisjećanje. U dva su istraživanja ispitali koliko dobro ove tri metode

utječu na učenje vokabulara. U prvom su eksperimentu svakoga pojedinca poučili drugačijom metodom. Zadatak leksičke odluke s usmjeravanjem pokazao je da su se parovi u kojima je usmjerivač bio povezan s ciljnom riječi brže prepoznavali od onih u kojima usmjerivač i ciljna riječ nisu bili povezani. Međutim, nisu otkrili nikakav učinak metode poučavanja. Stoga su u drugom eksperimentu svakoga ispitanika poučavali sa sve tri metode. Učinak metode poučavanja polučile su metoda poučavanja prevođenjem i metoda crno-bijelih slikovnih kartica. Nadalje prijevodni parovi koji su se poučavali pomoću metode prevođenjem brže su se prepoznavali nego oni koji su se poučavali metodom crno-bijelih slikovnih kartica. Tu pojavu objašnjavaju time što su se ispitanici više oslonili na pravopisnu sličnost između usmjerivača i ciljne riječi, dok je pri metodi poučavanja crno-bijelim slikovnim karticama potrebna semantička aktivacija. To je u skladu s revidiranim hijerarhijskim modelom koji postulira jače veze između leksičkih reprezentacija u početnim stadijima dvojezičnosti (Kroll i Stewart, 1994). Altarriba i Knickerbocker (2011) ističu da bi bilo dobro da se uskladi način poučavanja novih riječi i njihovo provjeravanje posebice u ranim razdobljima učenja J2. Također napominju da metoda poučavanja slikovnim karticama ima svoja ograničenja. Ona se očituje u poučavanju emocija. Međutim, taj bi se problem mogao riješiti tako da se u tom slučaju emocije poučava emotikonima koji su učestali u virtualnoj komunikaciji.

Povezano s istraživanjem koje su proveli Altarriba i Knickerbocker (2011) jest i ono koje su proveli Williams i Cheung (2011). U svojem su istraživanju pošli od toga da postoji epizodička i semantička memorija. Prva pohranjuje osobna iskustva, a druga znanje o svijetu. Kinesko-engleski dvojezični govornici učili su francuski kao J3 pri čemu im je engleski (J2) služio kao posrednik. Kineskim (J1) ciljnim riječima su u zadatku leksičke odluke s usmjeravanjem prethodile dvije vrste usmjerivača: ili francuski prijevodi kineske riječi ili asocijacije slabe asocijativne vrijednosti (primjerice, francuska je riječ *vjeverica* (*écureuil*) služila kao usmjerivač za njezin prijevodni ekvivalent na kineskom (松鼠) kao i za njezinu asocijaciju, *lješnjak* (果仁), na kineskom). Williams i Cheung polaze od toga da će cjelokupno značenje koje imaju pojedine riječi na engleskom kao J2 biti automatski preuzeto i preneseno na nove riječi na francuskom (J3). Otkrili su da do učinka usmjeravanja dolazi samo ukoliko su usmjerivač i ciljna riječ međusobni prijevodi. Zaključili su da novonaučene riječi na J3 ne preuzimaju sve semantičke informacije od J2. Isto tako nove riječi ne mogu polučiti učinak usmjeravanja čak i ako su asocijativno jače povezane. Učinak se usmjeravanja samo pojavljuje kod već poznatih francuskih riječi. U svojim su drugim eksperimentima poučavali

ispitanike nove riječi slikovnim karticama. Pokazalo se da riječi koje semantički i asocijativno nisu povezane mogu dovesti do bržega prepoznavanja ciljnih riječi ako su se pojavile na istoj slici tijekom faze poučavanja. Tako francuska riječ *prst* može pospješiti prepoznavanje ciljne riječi *leptir* na kineskom jeziku, samo zato što je na istoj slikovnoj kartici bio prikazan leptir kako sjedi na prstu. Williams i Cheung smatraju da je do toga došlo jer su ispitanici iskustveno doživjeli leptir i prst u istoj situaciji. Oni objašnjavaju tu pojavu pomoću modela memorije višestrukoga traga (*multiple-trace model*) (Hintzman, 1986). Prema tom modelu svako iskustvo stvara trag u pamćenju. Drugim riječima, pamćenje je sastavljeno od iskustvenih tragova. Kada se u zadatku leksičke odluke s usmjeravanjem prikaže usmjerivač, on istodobno ponovno aktivira tragove u pamćenju i semantička obilježja povezana s njime. Ta bi semantička obilježja usmjerivača aktivirala semantička obilježja u ciljnoj riječi što bi uzrokovalo usmjeravanje prijevodnoga ekvivalenta kao ciljne riječi. Budući da asocijativno povezana ciljna riječ nije bila prisutna u fazi poučavanja, tijekom rješavanja zadatka leksičke odluke s usmjeravanjem ne dolazi do usmjeravanja. Drugačije je pak u eksperimentima u kojima su se nove riječi poučavale slikovnim karticama. Obilježja riječi *prst* postala su dijelovi memorijskih tragova u fazama poučavanja nove riječi *leptir*. Kada se riječ *leptir* pojavila kao usmjerivač, aktivirala je uz svoje memorijske tragove i tragove riječi *prst*. Zbog toga je došlo do usmjeravanja. Stoga bi bilo najbolje da se nove riječi poučavaju pomoću slika jer se na taj način izbjegavaju krive pretpostavke o tome što te riječi zapravo znače. Slike i stvarni predmeti nisu toliko dvoznačni koliko riječi koje često imaju više značenja i ne pokrivaju ista značenja u oba jezika (Williams i Cheung, 2011).

De Groot i van Hell (2005) također su proučavale razne metode poučavanja stranoga jezika. Slično kao Altarriba i Knickerbocker (2011) zaključile su da se apstraktne riječi ne mogu prikazati slikama. Kako bi se govornicima olakšalo učenje novih stranih riječi, često se predlagala uporaba mnemotehnike ključnih riječi (*keyword method*) (De Groot i van Hell, 2005). Ta se metoda temelji na tome da se riječ na stranom jeziku koju se želi naučiti treba asocijativno povezati s drugom riječi na vlastitom jeziku koja zvuči slično. Tada osoba stvara interaktivnu predodžbu koja povezuje riječ na J1 s riječi na J2. De Groot i van Hell (2005) ističu da ta metoda može pospješiti učenje i pamćenje jer se koristi verbalnim i slikovnim sustavom ljudskoga pamćenja. To je u skladu s teorijom dvostrukoga kodiranja Allana Paivija (Paivio, 1969). De Groot i van Hell (2005) napominju isto tako da je niz istraživanja došao do zaključka da metoda ključnih riječi nije pogodna za dvojezične govornike koji su na

naprednijem stupnju poznavanja jezika (Moore i Surber, 1992) ili za višejezične govornike (van Hell i Candia Mahn, 1997). Uporabi te metode također se suprotstavljaju Barcroft, Sommers i Sunderman (2011). Oni napominju da se riječi koje su naučene pomoću metode ključnih riječi ne prepoznaju brže od onih koje su pojedinci učili napamet opetovanim ponavljanjem niti se u prepoznavanju prvih čini manje grešaka nego u prepoznavanju drugih. Riječi na J2 koje su se učile pomoću metode ključnih riječi utječu na sporije prepoznavanje ciljnih riječi u zadatku leksičke odluke s usmjeravanjem. Ako riječi koje su naučene pomoću metode ključnih riječi posluže kao usmjerivači, one odgađaju prepoznavanje jer se dodatno moraju semantički obraditi. Osim toga u svojem su istraživanju Barcroft, Sommers i Sunderman vidjeli da su se ispitanici nesrazmjerno češće oslanjali na asocijacije nastale metodom ključnih riječi kako bi pristupili mentalnom leksikonu od onih ispitanika koji su učili riječi napamet opetovanim ponavljanjem. Zaključili su to na temelju značajnih razlika u vremenu reakcije ispitanika na povezane i nepovezane parove usmjerivača i ciljnih riječi. Stoga se Barcroft, Sommers i Sunderman suprotstavljaju korištenju metode ključnih riječi u poučavanju neovisno o tome događa li se to unutar ili izvan učionice. Ističu da se metoda ključnih riječi može upotrijebiti kada se u vrlo kratkom vremenu treba naučiti veći broj riječi za određene prigode (primjerice kratak boravak u inozemstvu), a pojedinac ne želi steći esencijalno poznavanje J2.

Može se zaključiti da metode kojima se poučavaju nove riječi mogu mnogo utjecati na njihovo usvajanje. Naši su ispitanici bili poučavani različitim metodama tijekom svojega školovanja. Stoga nismo mogli utjecati na ovaj čimbenik.

13.2. Uloga kognata i konkretnih riječi

Važnu ulogu u poučavanju stranoga jezika igraju kognati (riječi koje imaju isto ili slično značenje, ortografiju i fonologiju u dva ili više jezika). Pokazalo se da kognati polučuju veći učinak kod dvojezičnih govornika koji još nemaju velikoga iskustva u J2, dok manji kod iskusnih dvojezičnih govornika (De Groot i Keijzer, 2000). Lotto i De Groot (1998) poučavali su ispitanike kognatima pomoću slika. Zaključile su kako slika aktivira riječ na J1 što ispitanicima olakšava prepoznavanje sličnosti između riječi na J1 i nove riječi na J2. Peters i Webb (2018) istražili su slučajno usvajanje novih riječi gledanjem televizijskoga programa. Proučavali su četiri čimbenika: frekvencija pojavljivanja riječi, kognatnost, prijašnje poznavanje J2 i relevantnost za razumijevanje sadržaja emisije. Otkrili su da osim potonjega

čimbenika ostala tri čimbenika utječu na uspješno prisjećanje značenja i prepoznavanje riječi na J2 kod nizozemsko-engleskih dvojezičnih govornika. Pokazalo se da kognati najviše utječu na učenje novih riječi. Učinak kognata bio je osam puta veći u prisjećanju riječi na J2 od nekognata, a 2.5 puta veći u prepoznavanju riječi na J2. Zaključili su da kognati imaju veći utjecaj tijekom slušanja nego pisanja. Razlog toj pojavi vide u tome što govornici ne mogu ponovno preslušati rečeno, pa se zbog toga moraju više oslanjati na sličnosti s J1 (De Wilde, Brysbaert i Eyckmans, 2019a). Comesaña, Soares i Lima (2010) u svojem su istraživanju ispitale rano učenje J2 kod djece s obzirom na metodu i vrstu riječi (kognati u odnosu prema nekognatima). Kod dvojezičnih govornika koji su tek krenuli s učenjem J2 semantičko je procesiranje posredovano leksičkom reprezentacijom putem J1, a kognati se jednostavnije uče jer imaju jaču leksičku vezu s J1. Time su potvrdile pretpostavke RHM-a. Van Hell i Dijkstra (2002) otkrili su da kod trojezičnih govornika (nizozemski kao J1, engleski kao J2 i francuski kao J3) koji su slabije vrsni na J3, kognati na J3 ne pospješuju prepoznavanje kognata na J1. Uočili su samo da kognati na J2 utječu na prepoznavanje kognata na J1. Drugim riječima, kognati se na J3 ne razlikuju od nekognata. Ako su govornici vrsniji na J3, tada se može uočiti da kognati na J3 pospješuju prepoznavanje riječi na J2 i J1. Prema tome, vrsnost govornika utječe na učinak kognata. Budući da su kognati vrlo bitni u početnom poučavanju njemačkoga jezika, može se pretpostaviti da su i naši ispitanici na taj način učili njemački jezik. Međutim, prošlo je već previše vremena kako bi se sjetili na koji su način učili svoje prve riječi na njemačkom jeziku.

De Groot i Keijzer (2000) te Lotto i De Groot (1997) uspoređivali su koliko se dobro ispitanici sjećaju riječi koje su učili nakon tjedan dana. Pokazalo se da one riječi koje se najlakše uče, a to su bili kognati i konkretne riječi, najdulje ostaju u sjećaju ispitanika, dok se nekognati i apstraktne riječi teže uče i brže zaboravljaju. Ellis i Beaton (1993) su dovele učinak konkretnosti u vezu s teorijom dvostrukoga kodiranja Allana Paivija (Paivio, 1969). Smatraju da konkretne riječi imaju verbalnu i perцепijsku reprezentaciju u ljudskom pamćenju. Zbog toga pojedinac može lakše pristupiti konkretnim riječima nego apstraktnima koje imaju samo verbalnu reprezentaciju. Upravo smo se zbog te činjenice u našim eksperimentima poslužili samo konkretnim riječima ne bismo li tako povećali učinak usmjeravanja. Lotto i de Groot (1997) ustanovile su nadalje da se visokofrekventne riječi prepoznaju brže nego niskofrekventne riječi. De Wilde, Brysbaert i Eyckmans (2019a) također su utvrdili veći učinak frekvencije na J2 za vrsnije dvojezične govornike. Što su

govornici vrsniji, znaju više visokofrekventnih riječi jer su se češće susretali s njima. Stoga Lotto i de Groot (1997) te de Wilde, Brysbaert i Eyckmans (2019a) predlažu da se na početku učenja novoga jezika vodi računa o tome koji će se tekstovi pojaviti u udžbenicima stranoga jezika. Ondje se trebaju naći tekstovi s mnogo visokofrekventnih riječi. Na taj se način početnim dvojezičnim govornicima omogućava da imaju pristup većem broju tekstova, posebice onima koji se ne nalaze u udžbenicima, tj. onima s kojima se susreću u svakodnevnom životu šecući ulicama (npr. natpisi na plakatima, novinski članci i sl.). Drugim riječima, u udžbeničkim tekstovima ne bi se trebalo pojaviti mnogo specifičnih riječi vezanih uz određeno tematsko područje, već više kognata i visokofrekventnih riječi jer će se tako brzo povećati vokabular na J2 kod početnih dvojezičnih govornika. Polazeći od te tvrdnje, pri generiranju ispitnoga materijal oslonili smo se na one riječi koje se uče na početnim stupnjevima J2.

13.3. Uloga distribucijskih semantičkih modela

Distribucijski semantički modeli poput *word2vec*-a mogu poslužiti kako bi se učenicima olakšalo učenje J2. Materijal za poučavanje J2 valja odabrati tako da se u početnim fazama često pojavljuju visokofrekventne riječi. To znači da bi se DSM-ovima odredile riječi koje su određene frekvencije. Njima bi se nadalje odredile riječi koje su na određenoj značenjskoj udaljenosti od njih. Drugim riječima, učenicima se na taj način može preporučiti koji im tekst najbolje odgovara kako bi proširili svoje poznavanje stranoga jezika. Slično su učinili Franceschetti i sur. (2001) za tekstove koje obrađuju tematiku iz fizike.

Wolfe i sur. (1998) dali su ispitanicima tekstove o načinu funkcioniranja ljudskoga srca različitoga stupnja složenosti. Proveli su testove razumijevanja i općega znanja neposredno prije i nakon što su ispitanici pročitali tekst. Zaključili su da LSA može potvrditi da ispitanici najviše znanja usvajaju ako je tekst optimalne težine. Na taj se način pokazalo kako se LSA može upotrijebiti za procjenu znanja. Isto se tako LSA-om može koristiti kako bi se svakom učeniku dao tekst koji po svojoj težini i složenosti najbolje odgovara njegovu trenutačnom znanju (Rehder i sur., 1998). Nadalje se upotrebljavaju modeli poput *word2vec*-a za analizu sažetaka jer je to vrlo uspješan način učenja J2 (Crossley i sur., 2019; Ryu, 2020).

Našim smo eksperimentima pokazali kako ispitni materijal generiran odabirom ispitivača i pomoću računalnoga modela *word2vec* polučuje podjednake rezultate. Vrsnost se kod učenika može povećati tako da im se više ne preporučuju tekstovi za daljnje učenje

nasumično, već da se koristi računalnim modelima koji mogu odrediti tekstove koje najbolje odgovaraju učeniku.

13.4. Uloga dobi i okruženja

Valja uzeti u obzir i okruženje u kojem se uči J2 te dob. Učenje se razlikuje između male djece, mladih i odraslih. Dob je čimbenik na koji nismo mogli utjecati prilikom odabira kandidata za naše istraživanje zbog ograničenja rada institucija. Naši su ispitanici uglavnom nastavnici koji su učili njemački jezik u školskom okruženju (61.11% i 54.55%). No u našim su eksperimentima pokazali veliku vrsnost što se potvrđuje i anketom samoprocjene.

Muñoz (2008) je ispitala utjecaj dobi na učenje J2. Analizirajući relevantnu literaturu, ustanovila je da mlađa djeca radije uče dodirom i pokretnom. Kako rastu, sve se više okreću auditivom ili vizualnom načinu učenja. Koji će pak prevagnuti, ovisi o kulturi i okolini u kojoj se nalaze. To je vjerojatno i razlog zašto su u istraživanju Lotto i de Groot (1998) prijevodni ekvivalenti imali veći učinak nego asociranje riječi sa slikama ili predmetima. Muñoz (2008) je nadalje uočila da su učenici u pubertetu manje motivirani u uspoređi sa starijim adolescentima. Isto tako ustanovila je da su mlađa djeca oduševljena aktivnostima u učionici ukoliko su ispunjeni određeni uvjeti. U razrednom okruženju treba vladati dobra radna atmosfera, treba postojati pozitivan stav prema učenju, učenici trebaju uživati u aktivnostima koje se provode i treba postojati pozitivan stav učitelja prema poučavanju. Kod starijih učenika navedeni pak čimbenici nisu toliko bitni. Hoće li učiti neki jezik ili ne, više ovisi o tome vide li korist u njegovu znanju. Andraka i Jurković (2015) istražile su aktivnosti za učenje vokabulara u udžbenicima ranoga učenja engleskoga jezika kao J2 u Hrvatskoj. Zaključile su da autori udžbenika prepuštaju odluku o strategijama poučavanja učiteljima. Stoga se može pretpostaviti da o stavu učitelja i njihovoj stručnosti ovisi koliko i kako će učenici učiti riječi na J2.

DeKeyser (2000) otkrio je da nijedna odrasla osoba ne može postići razinu izvornoga govornika ako nije poučavana eksplicitno i analitički pristupila J2. Drugim riječima, odrasli govornici nisu u stanju da usvoje nov jezik na isti način kako to čine djeca. Odrasli govornici pristupaju J2 tako što ga raščlanjuju i analiziraju njegovu strukturu. Ističe nadalje kako je kod odraslih govornika ključna njihova nadarenost koja kod djece ne igra veliku ulogu jer djeca uče J2 implicitno. Međutim, takav način usvajanja J2 pretpostavlja veliku količinu jezičnoga materijala koji se mora pružiti maloj djeci. Kako bi se to postiglo, djeca moraju biti u

okruženju u kojem se neprestano govori J2, a ne u razrednom okruženju. Prema trenutačnom su kurikulu djeca izložena J2 najčešće dva ili u višim razredima tri školska sata tjedno. Zbog toga što su učenici vrlo malo izloženi stranom jeziku, Holderbaum i Fumagalli de Salles (2011) ističu da se djeca više pouzdaju u kontekst kako bi brže pristupila mentalnom leksikonu jer još nisu dovoljno ovladala vještinom čitanja.

Osobnim smo razgovorom s ispitanicima došli do zaključka da je metoda kojom su najbolje učili njemački jezik metoda razgovora uz povremeno eksplicitno objašnjavanje gramatičkih zakonitosti. Međutim, kako smo već istaknuli, nismo mogli utjecati na način poučavanja naših ispitanika.

13.5. Utjecaj zadatka, aktivnosti i prestiža

Sunderman (2011) je pomoću Deese-Roedinger-McDermottove paradigme (DRM) zaključila da je najteži zadatak za dvojezičnoga govornika povezati određenu riječ s odgovarajućim konceptom i izreći ju. Paradigmom DRM-a istražuju se lažna sjećanja tako da se ispitanicima daje popis riječi koji trebaju zapamtiti. U drugom im se koraku daje nov popis riječi koji se sastoji od riječi koje su bile na prethodnom popisu i kritičke riječi. Primjerice, ako su na prvom popisu bile riječi *bolnica*, *lijek*, *medicinska sestra*, osoba bi se mogla dosjetiti i riječi *liječnik* bez obzira na to što ona nije bila ondje navedena. Sunderman je otkrila da ispitanici najviše lažnih sjećanja imaju na J1 te da vrsniji dvojezični govornici imaju više lažnih sjećanja na J2 od onih koji su manje vrsni. Prema tome, manje vrsni govornici s manjom bi vjerojatnošću aktivirali semantički asociranu riječ na J2. Ta pojava potvrđuje da manje vrsni govornici nisu u stanju izravno pristupiti konceptualnoj reprezentaciji za razliku od vrsnijih govornika koji to mogu (van Hell i Tanner, 2012). Takvo je objašnjenje u skladu s našim rezultatima koji su pokazali da naši dvojezični govornici mogu pristupiti konceptima i tako aktivirati asocijacije u oba jezika.

Povezivanje koncepta i riječi na J2 vrlo je teško na nižim stupnjevima poznavanja jezika jer govornici pokazuju poteškoće s inhibiranjem J1. Zbog toga bi se u nastavi valjalo više usredotočiti na produktivne vještine kao što su govorenje i pisanje (Sunderman, 2011; de Wilde, Brysbaert i Eyckmans, 2019a; de Wilde, Brysbaert i Eyckmans, 2019b). Sunderman (2011) tako predlaže da se u nastavi stranih jezika koristi aktivnostima poput opisivanja predmeta pri čemu su određene riječi zabranjene (slično igri *Taboo*), igranjem uloga (*role play*) ili ekstenzivnim čitanjem stripova i novina na stranom jeziku. U svojem su pak

istraživanju de Wilde, Brysbaert i Eyckmans (2019b) utvrdili da aktivnosti poput čitanja, slušanja glazbe ili gledanja televizijskih filmova ili serija na J2 uz podslove manje pridonose povećanju vrsnosti kod djece od onih aktivnosti koje se temelje na interakciji s drugim ljudima poput rasprave na društvenim mrežama, igranju računalnih igrica te razgovora na J2. Drugima riječima, pasivna uporaba jezika ne pridonosi tako puno razvoju vrsnosti kod mladih dvojezičnih govornika koji su početnici za razliku od aktivne uporabe J2. Isto tako smatraju kako drugi čimbenici poput inteligencije i nadarenosti za jezike također mogu igrati stanovitu ulogu u povećanju vrsnosti. Sudionici su našega drugog eksperimenta u anketi istaknuli da je njihovoj vrsnosti u poznavanju njemačkoga jezika najviše doprinijelo gledanje televizije i čitanje. Sličnoga su stava i ispitanici u prvom eksperimentu. Prema njima najviše je čitanje i učenje, a tek onda gledanje televizije doprinijelo poznavanju njemačkoga jezika. Međutim, valja napomenuti da je gotovo polovica ispitanika u oba eksperimenta bar jednom bila više od pola godine u zemljama njemačkoga govornog područja (1. eksperiment: 18, 2. eksperiment: 25).

De Wilde, Brysbaert i Eyckmans (2019b) zaključuju da u sredinama u kojima se engleski kao J2 smatra prestižnim jezikom i u kojima ima mnogo prilika da se s njim susreće izvaninstitucionalno na različitim aktivnostima dolazi do neformalnoga učenja i do usvajanja J2. Sličnoga su stava Mihaljević Djigunović i Geld (2002) koje su istražile mogućnosti kontekstnoga učenja engleskoga jezika (Elgort i sur., 2018). Engleski je jezik u Hrvatskoj jezik prestiža. Mogućnosti se za neformalno učenje engleskoga jezika nalaze posvuda (filmovi, serije, knjige, internet, zabavna kultura, promidžba i sl.) što je dovelo do toga da ima posebno mjesto među stranim jezicima. To se također očituje i u mnogobrojnim posuđenicama iz engleskoga jezika. Zaključile su da su dob, razina obrazovanja, znanje o svijetu i znanje o engleskom, duljina učenja engleskoga, interes, motivacija, frekvencija riječi, izgovorljivost i dostupnost čimbenici koji utječu na učenje riječi na J2.

13.6. Utjecaj udaljenosti između jezika

Njemački i hrvatski jezik na prvi se pogled smatraju udaljenim jezicima. Prvi pripada germanskoj skupini jezika, dok drugi slavenskoj. Pandžić (2011) je u svojem radu prikazao razlike u morfologiji između tih dvaju jezika. Međutim, otkrio je i neke strukturne sličnosti kao što su postojanje triju rodova, deklinacija pridjeva, slaba razvijenost složenica i sl. Slično je tome Svodoba (2018), analizirajući sintaktičke konstrukcije u njemačkom i hrvatskom

jeziku, zaključila da oba jezika dijele vrste rečenica. Budući da se većina ispitanika profesionalno bavi njemačkim jezikom i gotovo su mu svakodnevno izloženi, pretpostavljamo da je možda i implicitno poznavanje tih sličnosti pridonijelo njihovoj vrsnosti.

Schepens, van der Silk i van Hout (2015) proveli su istraživanje s višejezičnim govornicima kojima je nizozemski J3. Istražili su kakav utjecaj ima učinak udaljenosti između J1 i J2 na J3. Uočili su da na lakoću učenja utječu leksička i morfološka udaljenost između pojedinih jezika. Otkrili su da manja udaljenost između J2 i J3 utječe na to da se lakše može naučiti J3. J3 će se teže naučiti ako su J1 i J2 leksički udaljeniji i ako im je morfološka kompleksnost manja. Do sličnoga su zaključka došle Flynn, Foley i Vinnitskaya (2004) predstavljajući svoj model kumulativnoga poboljšanja (*cumulative-enhancement model*). Uspoređujući obrazac usvajanja i tvorbe odnosnih rečenice kod djece i odraslih, otkrile su da J1 nema povlašten status tijekom učenja drugih jezika. Pokazalo se nadalje da strukture koje postoje u jednom jeziku mogu utjecati na njihovo brže učenje u drugom jeziku te da prijašnje znanje J2 pozitivno utječe na J3. Prema tome poznavanje J2 više utječe na učenje J3 nego pojava da pojedinac prije toga uopće nije usvojio J2. De Groot i van Hell (2005) ističu da će se strukture i značenja više razlikovati s povećanjem udaljenosti između J1 i J2. Posljedica će toga biti otežano povezivanje J1 i J2.

13.7. Utjecaj fonologije na učenje J2

Njemački i hrvatski jezik u velikoj mjeri imaju glasove koji se slično izgovaraju. Iznimke mogu biti glasovi poput *r* ili *h* koji se ponešto drugačije izgovaraju u tim dvama jezicima (Ščukanec, 2008; Šnjarić, 2013). Budući da su sudionici naših eksperimenata uglavnom učitelji i nastavnici njemačkoga jezika, ovladali su pravilima izgovora i primjereno oblikuju sve glasove u oba jezika. Osim toga gotovo je polovica ispitanika u oba eksperimenta više od šest mjeseci živjela u nekoj od zemalja u kojima se govori njemačkim jezikom (18 odnosno 25).

Dvojezični govornici koji se nalaze na govornom području J2 i uključeni su u svakodnevni život te zajednice, ubrzavaju usvajanje J2. Pojačanom uporabom J2 dolazi do inhibicije J1. Takvom se inhibicijom J1 može potaknuti jačanje leksičko-konceptualnih veza jer se sprečavaju izravne leksičke veze između J2 i J1 (Linck, Kroll i Sunderman, 2009). Budući da je takvo učenje za većinu ljudi teško izvedivo, Jeong i sur. (2010) predlažu da se J2 ne uči na temelju tekstova (*text-based learning*), već tako da se simuliraju situacije koje se

mogu očekivati u stvarnom životu (*situation-based learning*). Mjerenjem aktivnosti mozga ustanovili su da bi bilo najbolje ako bi način učenja J2 zrcalio onaj J1, a to je u okruženju koji odgovara stvarnom životu.

Bitnu ulogu u određivanju vrsnosti na J2 ima izgovor (Lotto i de Groot, 1998; de Groot i van Hell, 2005, Trofimovich, 2011). Ako su fonotaktička pravila na J2 slična onima na J1, nove će se riječi brže naučiti. Primjerice, Trofimovich (2011) je istražio percipiraju li francusko-engleski dvojezični govornici u Kanadi razliku između /ð/ i /d/ odnosno /θ/ i /t/ te mogu li pravilno izgovoriti /ð/ i /θ/. Predlaže da bi učenike trebalo slikovnim pričama poticati na to da vježbaju izgovor. Smatra da bi se izgovor nekoga glasa mogao lakše povezati sa značenjem riječi gledanjem slike nekoga predmeta i slušanjem kako ga se izgovara. Isto tako ističe da bi poučavanje veza između pravopisa i pravogovora pozitivno djelovalo na govornike u izgradnji leksikona na J2 jer će na taj način moći predvidjeti kako se riječi izgovaraju s obzirom na njihov pisani oblik. Takav način poučavanja pogodan je za njemački jezik jer su veze između toga kako su riječi napisane i kako se trebaju izgovarati jasne.

Bialystok (2007) ističe da je fonološka svjesnost važna za razvoj pismenosti kod djece. Rezultati njezina istraživanja, naime, upućuju na to da se čitalačka vještina brže razvija kod djece s osnovnom fonološkom svjesnosti bez obzira na to na kojem je jeziku započeto poučavanje pismenosti. Fonološka se svjesnost koja se stekla na jednom jeziku može prenijeti na sposobnost čitanja na drugom jeziku. Zaključno se može reći da razvoj fonološke svjesnosti ovisi o tome koje jezike dijete govori, fonološkim vezama između tih jezika, zadacima kojima se provjerava fonološka svjesnost i iskustvu koje dijete posjeduje u stjecanju pismenosti.

13.8. Posljedice dvojezičnosti

Bialystok (2005; 2007) je detektirala mnoge posljedice koje usvajanje drugoga jezika može imati na dvojezično dijete. Ustanovila je da se kod mlađih govornika vrsnost u usmenom izražavanju i čitanju teže razvija jer se mora razviti za svaki jezik posebno. Zbog toga se učinak dvojezičnosti na vrsnost u usmenom izražavanju može smatrati negativnim. Isto se tako vidjelo da dvojezična djeca kasne u početnom razvoju vokabulara i vrsnosti usmenoga izražavanja za jednojezičnom djecom. Nadalje dvojezična djeca pokazuju da su sporija u rješavanju matematičkih zadataka riječima od aritmetičkih zadataka. Do toga dolazi jer dvojezična djeca još nisu toliko vrsna na J2 kako bi razumjela postavljen zadatak. Isto se

tako pokazalo da neki odrasli dvojezični govornici nešto sporije rješavaju aritmetičke zadatke. Do te pojave dolazi kada su zadaci postavljeni na jeziku kojim slabije vladaju. Dvojezične osobe su se, međutim, pokazale kreativnijima i stvari mogu sagledati iz drugoga kuta (Peal i Lambert, 1962). Osim toga dvojezičnost ubrzava razvoj kognitivnih funkcija povezanih s pozornosti i inhibicijom. Bialystok i Senman (2004) su u istraživanju s četverogodišnjim jednojezičnim i dvojezičnim govornicima u zadatku razlikovanja pojavnosti i realnosti zaključile da vrlo mladi dvojezični govornici mogu lakše suspregnuti informacije koje navode na pogrešan zaključak i kontrolirati svoju pozornost. U tom su zadatku, naime, četverogodišnjaci trebali odgovoriti na tri pitanja vezana uz određen predmet. Na prvi se pogled činilo da stvar koju vide predstavlja jedan predmet. Međutim, detaljnijim gledanjem ispostavilo se da se radi o sasvim drugom predmetu. Primjerice, na prvi se pogled određen predmet činio plastičnim snjegovićem, no ispostavilo se da se zapravo radi o knjizi. Prva dva pitanja odnosila su se na pojavnost predmeta kada su ga djeca prvi put uočila (što si mislio kada si prvi put vidio predmet?), a treće se odnosilo na realnost predmeta, odnosno na ono što on predstavlja u stvarnosti (što je to zapravo?). Budući da pojavnost predmeta ne odgovara njegovoj realnosti, djeca aktivno moraju potisnuti pojavnost i odrediti što predmet zaista predstavlja susprežući vanjštinu koja navodi na krivi zaključak. Dvojezična su djeca u tom zadatku nadmašila jednojezičnu. Budući da se jezični sukob između J1 i J2 kod dvojezičnih govornika odvija neprestano i neovisno o dobi, on pojačava sustav izvršne kontrole štiteći ga istodobno od propadanja (Bialystok, Craik i Luk, 2008).

Iako su u različitim istraživanjima uočene poteškoće s kojima se dvojezični govornici moraju suočavati, naši su ispitanici sve takve možebitne poteškoće savladali do te mjere da su u eksperimentima u kojima su sudjelovali pokazali odlike onih ispitanika koji su veoma vrsni na svojem J2.

13.9. Preporuke za poučavanje J2

Na temelju našega istraživanja, ali i prikazanih istraživanja, iznijet ćemo u nastavku niz od 12 preporuka kojima se ubuduće može voditi pri poučavanju i oblikovanju didaktičkoga materijala. Na taj bi se način učenicima stranoga jezika moglo olakšati stjecanje novoga vokabulara i ubrzati ga. Valja, međutim, napomenuti da se u našem okruženju vrlo rijetko upotrebljavaju računalni modeli u nastavi. Stoga bi bilo dobro da se za početak učitelje i nastavnike upozna s mogućnostima koje pružaju računalni modeli u obradi podataka,

odabiru primjerenoga teksta koji optimalno odgovara trenutačnom znanju učenika i vrednovanju učeničkih uradaka. Nastava na daljinu potaknula je učitelje i nastavnike da se pojačano koriste digitalnim alatima. Također se mnogo radi na osuvremenjivanju nastavnih metoda za digitalno okruženje. Stoga bi u tom smislu valjalo uključiti distribucijske semantičke modele i skrenuti pozornost na mogućnosti koje oni pružaju u poboljšanju nastave.

1. U ranim razdobljima učenja J2 treba uskladiti način poučavanja novih riječi i njihovo provjeravanje (Altarriba i Knickerbocker, 2011).
2. Nove riječi treba poučavati pomoću slika i stvarnih predmeta jer nisu toliko dvoznačni (Williams i Cheung, 2011).
3. Učenike treba izričito poučavati vezama između pravopisa i pravogovora jer će na taj način moći predvidjeti kako se riječi izgovaraju s obzirom na njihov pisani oblik (Trofimovich, 2011). Isto tako gledanjem slike nekoga predmeta i slušanjem kako se izgovara može se povezati pravilan izgovor sa značenjem riječi (Trofimovich, 2011).
4. Treba izbjegavati metodu ključnih riječi u poučavanju riječi na J2 (Barcroft, Sommers i Sunderman, 2011).
5. U udžbeničkim bi se tekstovima trebalo pojaviti više kognata i visokofrekventnih riječi jer će se tako brzo povećati vokabular početnih dvojezičnih govornika na J2 (de Wilde, Brysbaert i Eyckmans, 2019a).
6. Materijal za poučavanje J2 valja oblikovati pomoću DSM-ova tako da se za odabrane visokofrekventne riječi nađu druge riječi s kojima se često usporedno pojavljuju.
7. DSM-ovima se može koristiti kako bi se učenike uputilo na materijal koji je njima zanimljiv i koji pomaže povećanju vrsnosti na J2.
8. Kod mlađih učenika u razrednom okruženju treba vladati dobra radna atmosfera, treba postojati pozitivan stav prema učenju, učenici trebaju uživati u aktivnostima koje se provode i treba postojati pozitivan stav učitelja prema poučavanju (Muñoz, 2008).
9. Kod starijih učenika treba isticati važnost i praktičnu korist koju mogu imati učenjem J2 (Muñoz, 2008).
10. Češće se trebaju provoditi aktivnosti koje potiču razvoj produktivnih sposobnosti učenika (de Wilde, Brysbaert i Eyckmans, 2019b).
11. Učenike treba poticati da se izvan škole koriste stranim jezikom kroz razne aktivnosti (čitanje, gledanje televizijskih filmova i serija na J2 uz podslove, vođenje rasprava na

društvenim mrežama, igranje računalnih igrica sa suigračima iz stranih zemalja, razgovori s ljudima iz stranih zemalja) (de Wilde, Brysbaert i Eyckmans, 2019a; de Wilde, Brysbaert i Eyckmans, 2019b).

12. J2 se treba učiti u situacijama koje zrcale stvarne situacije (Jeong i sur., 2010).

14. ZAKLJUČAK

Cilj je ovoga istraživanja bio odrediti mogu li distribucijski semantički modeli (DSM) odražavati funkcioniranje mentalnoga leksikona kod ljudi te ako mogu, kako se ta činjenica može iskoristiti u oblikovanju didaktičkih materijala. To smo pokušali ustanoviti dvama eksperimentima. U prvom je eksperimentu ispitni materijal generiran pomoću diskretnih slobodnih asocijacija, a u drugom pomoću predviđačkoga modela DSM-a *word2vec*.

U oba je eksperimenta potvrđeno da se ciljne riječi koje su asocijativno povezane s usmjerivačima brže prepoznaju nego one koje to nisu. Ova je pojava u skladu s dvojezičnim interaktivnim aktivacijskim modelom BIA+. Povezani su parovi bliži u semantičkoj mreži i zbog toga se brže aktiviraju. Eksperimentima se pokazalo da naši ispitanici brže prepoznaju ciljne riječi na J1 ako su im prethodili usmjerivači na J2. Može se pretpostaviti da je razlog tomu vrsnost naših ispitanika na J2. Neujednačeni dvojezični govornici imaju bogatiji leksikon na J1 zbog čega se brže aktiviraju čvorovi u tom jeziku. Posljedica će toga toga biti ta da se ciljne riječi na J2 brže prepoznaju ako im je prethodio usmjerivač na J1. Za razliku od neujednačenih dvojezičnih govornika, kod ujednačenih dvojezičnih govornika aktivira se podjednak broj bliskih čvorova u oba jezika. Međutim, na brzinu prepoznavanja ciljnih riječi utječu mnogobrojni čimbenici. Primjerice, konkretne imenice, kojima se koristilo u ovom istraživanju, prepoznaju se brže nego apstraktne imenice. Na brže prepoznavanje ciljnih riječi na J1 može utjecati i to kako je zadatak postavljen te kako je sastavljen ispitani materijal. Vrsnost dvojezičnih govornika također utječe na brzinu prepoznavanja. Naši su ispitanici vrlo vrsni dvojezični govornici njemačkoga jezika i svakodnevno se služe njime. Osim toga većina je boravila na njemačkom govornom području. Ta činjenica može utjecati na brži pristup leksikonu i bolje povezivanje riječi s konceptima. Posljedica toga je povećanje simetrije pri usmjeravanju. Povećana vrsnost u jeziku utječe također na aktivaciju ciljnoga jezika i inhibiciju neciljnoga jezika te povećava kontrolu pozornosti i kognitivnu kontrolu. Prema tome, može se pretpostaviti da ispitanici koji su sudjelovali u našem istraživanju imaju izravan pristup konceptima. To bi značilo da su naši rezultati također u skladu s nekim postavkama revidiranoga hijerarhijskog modela (RHM). Razvojna komponenta modela polazi od toga da dvojezični govornik s povećanjem vrsnosti na J2 može izravno pristupiti konceptima. Prestaje im, dakle, pristupati preko leksičkih veza s J1. Kada će dvojezični govornik postati vrstan govornik J2, još se treba istražiti.

Pokazalo se da ne postoje razlike između ispitanika kod kojih su asocijativno-semantički povezani parovi usmjerivača i ciljnih riječi bili generirani pomoću diskretnih asocijacija i pomoću modela *word2vec*. Na temelju toga može se zaključiti da ovaj DSM zrcali način ljudskoga rasuđivanja. To je u skladu s prijašnjim istraživanjima koja su zaključila da je *word2vec* psihološki utemeljeniji od drugih DSM-ova, posebice modela prebrojavanja poput LSA-a i HAL-a, jer se koristi modelom asocijativnoga učenja koji su predložili Rescorla i Wagner (1972).

Cilj je svake nastave da kod učenika postigne što veću razinu poznavanja J2 u četirima vještinama: slušanje, govorenje, čitanje i pisanje. Zaključili smo da se to može najbolje postići multimodalnim načinom poučavanja. Metode poučavanja također ovise o dobi učenika. Dok je kod mladih dvojezičnih govornika važna atmosfera, odnos prema učitelju i predmetu, raznovrsnost provedenih aktivnosti, uporaba konkretnih predmeta, kod starijih je dvojezičnih govornika važno analitičko raščlanjivanje jezika te isticanje važnosti i koristi koju imaju učenjem jezika. DSM-ovi poput *word2vec*-a ne samo da mogu poslužiti za provjeru poznavanja jezika već mogu pomoći i u odabiru tekstova kojima će se poučavati strani jezik.

Na kraju možemo zaključiti da je predviđački model *word2vec* model DSM-a koji može zrcaliti ljudsko rezoniranje. Takvo je rezoniranje u skladu s određenim postavkama modela BIA+ i RHM-a o načinu funkcioniranja mentalnoga leksikona. Budući da se rezultati u zadatku leksičke odluke s usmjeravanjem ne razlikuju, ispitni se materijal generiran modelom *word2vec* može upotrijebiti umjesto onoga koji je generiran uobičajenim načinom, tj. odabirom ispitivača.

15. LITERATURA

Allan, Keit (2015). A history of semantics. Nick Riemer, ur. *The Routledge Handbook of Semantics*. London, New York: Routledge, 48-68.

Altarriba, Jeanette i Hugh Knickerbocker (2011). Acquiring second language vocabulary through the use of images and words. Pavel Trofimovich i Kim McDonough, ur. *Applying priming methods to L2 learning, teaching and reseach: Insights from Psycholinguistics*. Amsterdam: Benjamins, 21-48.

Andraka, Marija i Ankica Jurković (2015). Potiču li udžbenici za rano učenje stranog jezika razvoj strategija učenja vokabulara? *Strani jezici* 44(1): 29-53.

Andrews, Mark, Gabriella Vigliocco i David Vinson (2009). Integrating Experiential and Distributional Data to Learn Semantic Representations. *Psychological Review* 116(3): 463-498.

Andrews, Mark, Gabriella Vigliocco i David Vinson (2009). Integrating Experiential and Distributional Data to Learn Semantic Representations. *Psychological Review* 116(3): 436-498.

Anić i sur. (2003). *Hrvatski enciklopedijski rječnik*. Zagreb: Novi Liber.

Asr, Fatemeh Torabi, Jon A. Willits i Michael N. Jones (2016). Comparing Predictive and Co-occurrence Based Models of Lexical Semantics Trained od Child-directed Speech. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Barcroft, Joe, Mitchell S. Sommers i Gretchen L. Sunderman (2011). Some costs of fooling Mother Nature. U: Pavel Trofimovich i Kim McDonough, ur. *Applying priming methods to L2 learning, teaching and reseach: Insights from Psycholinguistics*. Amsterdam: Benjamins, 49-72.

Baroni, Marco i Alesandro Lenci (2010). Distributional Memory: A General Framework for Corpus-Based Semantics. *Association for Computational Linguistics* 36(4): 673-721.

Baroni, Marco i Roberto Zamparelli (2010). Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. Hand Li i Lluís Marquez, ur. *EMNLP '10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, Massachusetts: Association for Computational Linguistics, 1183-1193.

Baroni, Marco, Brian Murphy, Eduard Barbu, Massimo Poesio (2010). Strudel: A Corpus-Based Semantic Model Based on Properties and Types. *Cognitive Science* 34(2): 222-254.

Baroni, Marco, Geogiana Dinu, Germán Kruszewski (2014). Don't count, predict! A systemic comparison of context-counting vs. context-predicting semantic vectors. Kristina Toutanova i Hau Wu, ur. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, 238-247.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi i Eros Zanchetta (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43: 209-226.

Barsalou, Lawrence W. (1999). Perceptual symbol systems. *Behavioral and brain science* 22: 577-660.

Basnight-Brown, Dana M. (2014). Models of Lexical Access and Bilingualism. Robnerto R. Heredia i Jeanette Altarriba, ur. *Foundations of Bilingual Memory*. Springer: New York, 85-107.

Basnight-Brown, Dana M. i Jeanette Altarriba (2007). Differences in semantic and translation priming across languages: The role of language direction and language dominance. *Memory & Cognition* 35: 953-965.

Basnight-Brown, Dana M. i Jeanette Altarriba (2015). Multiple Translations in Bilingual Memory: Processing Differences Across Concrete, Abstract, and Emotion Words. *Journal of Psycholinguistic Research* 45(5): 1219-1245.

Berken, Jonathan A., Vincent L. Gracco, Jen-Kai Chen, Kate E. Watkins, Shari Baum, Megan Callahan i Denise Klein (2015). Neural activation in speech production and reading aloud in native and non-native languages. *NeuroImage* 112: 208-217.

Bialystok, Ellen i Lili Senman (2004). Executive Processes in Appearance-Reality Tasks: The Role of Inhibition of Attention and Symbolic Representation. *Child Development* 75(2): 562-579.

Bialystok, Ellen, Fergus Craik i Gigi Luk (2008). Cognitive Control and Lexical Access in Younger and Older Bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34(4): 859-873.

Bialystok, Ellen (2005). Consequences of Bilingualism for Cognitive Development. Judith F. Kroll i Annette M. B. de Groot, ur. *Handbook of Bilingualism: Psycholinguistic approaches*. New York: Oxford University Press, 417-432.

Bialystok, Ellen (2007). Acquisition of Literacy in Bilingual Children: A Framework for Research. *Language Learning* 57: 45-77.

Bisk, Yonatan, Ari Holtzman, Jesse Thomasson, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto i Joseph Turian (2020). Experience Grounds Language. arXiv:2004.10151v3 [cs.CL].

Bloomfield, Leonard (1933). *Language*. New York: Henry Hold and Co.

Bogdanović, Danijel (2015). *Noćni vlak za Dukku*. Zagreb: Hangar 7.

Bogunović, Irena i Bojana Ćoso (2019). Leksički pristup hrvatskih govornika engleskog s različitim razinama jezičnog znanja. *Suvremena lingvistika* 87: 1-22.

Borko, Harold i Myrna D. Bernick (1963). Automatic document classification. *Journal of the ACM* 10(3): 151-162.

Brender, Emily M. i Alexander Koller (2020). Climbing towards NLU. On Meaning, Form, and Understanding in the Age of Data. Dan Jurafsky, Joyce Chai, Natalie Schuler i Joel Tetreault, ur. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5185-5198.

Brenders, Pascal, Janet G. van Hell i Ton Dijkstra (2011). Word recognition in child second language learners: Evidence from cognates and false friends. *Journal of Experimental Child Psychology* 109: 383-396.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranac Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever i Dario Amodai (2020). Language Models are Few-Shot Learners. arXiv:2005.14165v4 [cs.CL].

Bruni, Elia, Nam Khanh Tran i Marco Baroni (2014). Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research* 49: 1-47.

Brysbaert, Marc i Andrew W. Ellis (2015). Aphasia and age of acquisition: are early-learned words more resilient? *Aphasiology* 30(11): 1240-1263.

Brysbaert, Marc i Boris New (2009). Moving beyond Kučera adn Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41(4): 977-990.

Brysbaert, Marc i Ton Dijkstra. Changing views on word recognition in bilinguals. José Morais i Géry d'Ydewalle, ur. *Bilingualism and second language acquisition*. Bruxelles: Royal Academies for Science and Art of Belgium.

Brysbaert, Marc, Eef Ameel i Gert Storms (2004). Semantic memory and bilingualism: a review of the literature and a new hypothesis. Roeberto Heredia, Jeanette Altarriba. *Foundations of Bilingual memory*. New York: Springer, 133-146

Brysbaert, Marc, Emmanuel Keuleers i Boris New (2011). Assessing the Usefulness of Google Books' Word Frequencies for Psycholinguistic Research on Word Processing. *Frontiers in Psychology* 2(27). DOI: 10.3389/fpsyg.2011.00027.

Brysbaert, Marc, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte i Andrea Böhl (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental psychology* 58(5): 412-424.

Brysbaert, Marc, Paweł Mandera i Emmanuel Keuleers (2017). Corpus linguistics. Annette M. B. de Groot i Peter Hagoort, ur. *Research methods in psycholinguistics and the neurobiology of language: a practical guide*. Oxford: Wiley.

Bullinaria, John A. i Joseph P. Levy (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods* 3: 510-526.

Bullinaria, John A. i Joseph P. Levy (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods* 44(3): 890-907.

Burgess, Curt i Kevin Lund (1997). Modelling Parsing Constraints with High-dimensional Context Space. *Language and Cognitive Processes* 12(2-3): 177-210.

Burgess, Curt i Kevin Lund (2000). The dynamics of meaning in memory. *Cognitive dynamics: Conceptual and representational change in humans and machines* 13: 17-56.

Burgess, Curt, Livesay, Kay i Kevin Lund (1998). Explorations in Context Space: Words, Sentences, Discourse. *Discourse Processes* 25(2-3): 211-257.

Burgess, Curt (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers* 30(2): 188-198.

Cergol, Kristina (2011). Jezična aktivacija i leksički pristup u stranom ili drugom i materinskom jeziku. Neobjavljena doktorska disertacija.

Chen, Yen-Chun, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng i Jingjing Liu (2020). UNITER: UNiversal Image-TExt Representation Learning. arXiv:1909.11740v3

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Chomsky, Noam (1986). *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger Publishers.

Chomsky, Noam (1995). Language and Nature. *Mind* 104(413): 1-61.

Coltheart, Max (1980). Iconic memory and visible persistence. *Perception & Psychophysics* 27: 183-228.

Comesaña, Montserrat, Ana Paula Soares i Cátia Lima (2010). Semantic representations of new cognate vs. noncognate words: Evidence from two second language learning methods. *Procedia - Social and Behavioral Sciences* 5: 199-203.

De Deyne, Simon, Amy Perfors i Daniel J. Navarro (2016). Predicting human similarity judgements with distributional models: The value of word associations. Yuji Matsumoto i Rashmi Prasad, ur. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan: The COLING 2016 Organizing Committee.

De Deyne, Simon, Danielle J. Navarro, Amy Perfors, Marc Brysbaert i Gert Storms (2018). The "Small World of Words" English word association norms for over 12.000 cue words. *Behavior Research Methods* 51: 987-1006. DOI 10.3758/s13428-018-1115-7.

De Deyne, Simon, Steven Verheyen i Gert Storms (2015). The role of corpus size and syntax in deriving lexico-semantic representations for a wide range of concepts. *Quarterly journal of experimental psychology* 68(8): 1643-4664.

De Groot, Annette M. B. (1992a). Determinants of Word Translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18(2): 1001-1018.

De Groot, Annette M. B. (1992b). Bilingual lexical representation: A closer look at conceptual representations. R. Frost i L. Katz, ur. *Orthography, phonology, morphology, and meaning*. Amsterdam: North-Holland, 389-412.

De Groot, Annette M. B. (2011). *Language and Cognition in Bilinguals and Multilinguals. An Introduction*. New York i Hove: Psychology Press.

De Groot, Annette M. B. i Gerard L. J. Nas (1991). Lexical Representation of Cognates and Noncognates in Compound Bilinguals. *Journal of Memory and Language* 30: 90-123.

De Groot, Annette M. B. i Janet G. van Hell (2005). The Learning of Foreign Language Vocabulary. Judith F. Kroll i Annette M. B. de Groot, ur. *Handbook of bilingualism: Psycholinguistic approaches*. New York: Oxford University Press, 9-29.

De Groot, Annette M. B. i Rineke Keijzer (2000). What Is Hard to Learn Is Easy to Forget: The Roles of Word Concreteness, Cognate Status, and Word Frequency in Foreign-Language Vocabulary Learning and Forgetting. *Language Learning* 50(1): 1-56.

De Groot., Annette B. M. (1993). Word type effects in bilingual processing tasks: Sort for mixed representational system. R. Schreuder i B. Weltens, ur. *The bilingual lexicon*. Amsterdam, Philadelphia: John Benjamin. 27-51.

De Saussure, Ferdinand (2000). *Tečaj opće lingvistike*. Zagreb: Institut za hrvatski jezik i jezikoslovlje.

De Wilde, Vanessa, Marc Brysbaert i June Eyckmans (2019a). Learning English Through Out-of-School Exposure: How Do Word-Related Variables and Proficiency Influence Receptive Vocabulary Learning? *Language Learning*. DOI: 10.1111/lang.12380.

De Wilde, Vanessa, Marc Brysbaert i June Eyckmans (2019b). Learning English through out-of-school exposure. Which levels of language proficiency are attained and which types of input are important? *Bilingualism: Language and Cognition* 1(15). DOI: 10.1017/S1366728918001062

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer i Richard Harshman (1990). Indexing by Latent Semantic Analysis. *Journal of the Association for Information Science and Technology* 41(6): 391-407.

DeKeyser, Robert M. (2000). The Robustness of Critical Period Effects in Second Language Acquisition. *Studies in Second Language Acquisition* 22: 499-533.

Demme, Silke, Hermann Funk i Christina Funk (2016). *Studio D AI*. Kurs- und Arbeitsbuch. Berlin: Cornelsen.

Dengler, Stefanie, Paul Rusch, Helen Schmitz i Tanja Sieber (2016). *Netzwerk AI*. Deutsch als Fremdsprache. München: Langenscheidt.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee i Kristina Toutanova (2019). BERT. Pre-training of Deep Bidirectional Transformers for Language Understanding. Jill Burstein, Christy Doran i Tamar Solorio, ur. *Proceedings of NAACL-HLT*. Minneapolis, Minnesota: Association for Computational Linguistics, 4171-4186.

Dijkstra, Ton (2003). Lexical Processing in Bilinguals and Multilinguals: The Word Selection Problem. Jasone Cenoz, Britta Hufeisen i Ulrike Jessner, ur. *The Multilingual Lexicon*. Springer: Dordrecht, 11-26.

Dijkstra, Ton i Walter J. B. van Heuven (1998). The BIA model and bilingual word recognition. Jonathan Grainger i Arthur M. Jacobs ur. *Localist Connectionist Approaches To Human Cognition*. Mahwah: Lawrence Erlbaum Associates, 189-225.

Dijkstra, Ton i Walter J. B. van Heuven (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition* 5: 175– 197.

Dijkstra, Ton i Walter J. B. van Heuven (2012). Word recognition in the bilingual brain. Miriam Faust, ur. *The handbook of neuropsychology of language*. West Sussex: Wiley-Blackwell, 449-471.

Dijkstra, Ton, Fermin Moscoso del Prado Martín, Béryl Schulpen, Robert Schreuder i R. Harald Baayen (2007). A roommate in cream: Morphological family size effects on interlingual homograph recognition. *Language and Cognitive Processes*. 20(1-2): 7-41.

Dijkstra, Ton, Jonathan Grainger i Walter J. B. van Heuven (1999). Recognition of Cognates and Interlingual Homographs: The Neglected Role of Phonology. *Journal of Memory and Language* 41: 496-518.

Dijkstra, Ton (2005). Bilingual Visual Word Recognition and Lexical Access. Judith F. Kroll i Annette M. B. de Groot, ur. *Handbook of bilingualism: Psycholinguistic approaches*. New York: Oxford University Press, 179-201.

Dimitropoulou, Maria, Jon Andoni Duñabeitia i Manuel Carreiras (2011). Masked translation priming effects with low proficient bilinguals. *Memory & Cognition* 39: 260-275.

Dumais, Susan T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers* 23(2): 229-236.

Duyck, Wouter i Marc Brysbaert (2004). Forward and backward number translation requires conceptual mediation in both balanced and unbalanced bilinguals. *Journal of Experimental Psychology: Human Perception & Performance* 30: 889-906.

Elgort, Irina, Marc Brysbaert, Michaël Stevens i Eva van Assche (2018). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition* 40(2): 341-366.

Ellis, Nick i Alan Beaton (1993). Factors Affecting the Learning of Foreign Language Vocabulary: Imagery Keyword Mediators and Phonological Short-Term Memory. *The Quarterly Journal of Experimental Psychology* 46A(3): 533-558.

Erdeljac, Vlasta. (2009). *Mentalni leksikon*. Zagreb: Ibis grafika.

Ettinger, Allyson i Tal Linzen (2016). Evaluating vector space models using human semantic priming results. Roi Reichart, Omer Levy, Felix Hill, Anna Korhonen, Kyunghyun Cho, Yoav Goldberg i Antoine Bordes, ur. *Proceedings of the 1st Workshop on Evaluating Vector Space Representation for NLP*. Berlin: Association for Computational Linguistics, 72-77.

Evert, Stefan i Alessandro Lenci (2009). Foundation of Distributional Semantic Models. ESSLLI '09. Bordeaux, 27.7.2009. predavanje, ppt.

Faaß, Gertrud i Kerstin Eckart (2013). SdeWaC - A Corpus of Parsable Sentences from the Web. Iryna Gurevych, Chris Biemann i Torsten Zesch. *Language Processing and Knowledge in the Web*. Berlin, Heidelberg: Springer, 61-68.

Feng, Yansong i Mirella Lapata (2010). Visual information in semantic representation. Ron Kaplan, Jill Burstein, Mary Harper i Gerald Penn, ur. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*. Los Angeles: Association for Computational Linguistics, 91-99.

Ferré, Pilar, Rosa Sánchez-Casas, Monserrat Comesaña i Josep Demetre (2016). Masked translation priming with cognates and noncognates: Is there an effect of words' concreteness? *Bilingualism: Language and Cognition*: 1-13.

Finkbeiner, Matthew, Kenneth I. Forster, Janet Nicol i Kumiko Nakamura (2004). The role of polysemy in masked semantic and translation priming. *Journal of Memory and Language* 51: 1-22.

Finkbeiner, Matthew (2005). Task-Dependent L2-L1 Translation Priming: An Investigation of the Separate Memory Systems Account. James Cohen, Kara T. McAllister, Kellie Rolstad i Jeff MacSwan, ur. *ISB4: Proceedings of the 4th International Symposium on Bilingualism*. Somerville: Cascadilla Press, 741-750.

Firth, John R. (1957). *Papres in Linguistics 1934-51*. London: Oxford University Press.

Flynn, Suzanne, Claire Foley i Inna Vinnitskaya (2004). The Cumulative-Enhancement Model for Language Acquisition: Comparing Adults' and Children's Patterns of Development

in First, Second and Third Language Acquisition of Relative Clauses. *International Journal of Multilingualism* 1(1): 3-16.

French, Robert M. (1998). A Simple Recurrent Network Model of Bilingual Memory. M. A. Gernsbacher i S. J. Derry, ur. *Proceedings of the 20th Annual Conference of the Cognitive Science Society*. Mahwah, New Jersey: Lawrence Erlbaum Associates, 368-373.

French, Robert M. i Maud Jacquet (2004). Understanding bilingual memory: models and data. *Trends in Cognitive Science* 8(2): 87-93.

Furnas, George W., Scott Deerwester, Susan T. Dumais, Thomas K. Landauer, Richard Harshman, Lynn A. Streeter i Karen E. Lochbaum (1988). Information Retrieval using a Singular Value Decomposition Model of Latent Semantic Structure. Chiaramella, Yves, ur. *Proceeding of the 11th annual international Association of Computing Machinery SIGIR conference on Research and development in information retrieval*. Grenoble, France, 465-480.

Gerard, Linda D. i Don L. Scarborough (1989). Language-Specific Lexical Access of Homographs by Bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15(2): 305-315.

Gimenes, Manuel i Boris New (2016). Wordlex: Twitter and blog word frequencies for 66 languages. *Behavior research methods* 48(3): 963-972.

Glenberg, Arthur M. i David A. Robertson (2000). Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning. *Journal of Memory and Language* 43(3): 379-401.

Goldberg, Yoav i Omer Levy (2014). *word2vec* Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *Computation and Language (cs.CL)*. arXiv:1402.3722 [cs.CV]

Goldman, Susan R., Thomas W. Hogaboam, Laura C. Bell i Charles A. Perfetti (1980). Short-Term Retention of Discourse During Reading. *Journal of Education and Psychology* 72(5): 647-655.

Golub, Gene H. i Charles F. van Loan (1996). *Matrix computations*. Baltimore i London: The John Hopkins University Press.

Grainger, Jonathan i Cheryl Frenck-Mestre (1998). Masked Priming by Translation Equivalents in Proficient Bilinguals. *Bilingualism. Language and Cognitive Processes* 13(6): 601-623.

Grainger, Jonathan, Katherine Midgley i Phillip J. Holcomb (2010). Re-thinking the bilingual interactive activation model from a developmental perspective (BIA-d). Michèle Kail i Maya Hickmann, ur. *Language Acquisition across Linguistic and Cognitive Systems*. Amsterdam: John Benjamins Publishing Company, 267-283.

Green, David W. (1993). Towards a Model of L2 Comprehension and Production. Robert Schreider i Bert Weltens, ur. *The Bilingual Lexicon*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Green, David W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Control* 1: 67-81.

Green, David W. (2000). Control, activation, and resource: a framework and a model for the control of speech in bilinguals. Li Wei, ur. *The Bilingual Reader*. London i New York: Routledge.

Green, David W. (2011). Biligual Words. Vivian Cook i Benedetta Bassetti, ur. *Language and Bilingual Cognition*. New York i Hove: Psychology Press, 229-240.

Grefenstette, Edward, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh i Marco Baroni (2013). Multi-Step Regression Learning for Compositional Distributional Semantics. Computation and Language. Alexander Koller i Katrin Erk, ur. *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*. Postdam, Germany: Associatoin for Computational Linguistics, 131-142.

Guasch, Marc, Rosa Sánchez-Casas, Pilar Ferré i José E. García-Albea (2011). Effects of the degree of meaning similarity on cross-language semantic priming in highly proficient bilinguals. *Journal of Cognitive Psychology* 23(8): 942-961.

Gulan, Tanja (2016). *Organizacija mentalnog leksikona kod bilingvalnih govornika*. Zagreb. Neobjavljena doktorska disertacija.

Günther, Fritz, Carolin Dudschig i Barbara Kaup (2015). LSAfun - An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods* 47: 930-944.

Günther, Fritz, Carolin Dudschig i Barbara Kaup (2016a). Latent semantic Analysis cosines as a cognitive similarity measure. Evidence from priming studies. *Quarterly Journal of Experimental Psychology* 69(4): 626-653.

Günther, Fritz, Carolin Dudschig i Barbara Kaup (2016b). Predicting Lexical Priming Effects from Distributional Semantic Similarities. A Replication with Extension. *Frontiers in Psychology* 7:16-46. DOI: 10.3389/fpsyg.2016.01646.

Günther, Fritz, Luca Rinaldi i Marko Marelli (2019). Vector-Space Models of Semantic Representation From a Cognitive Perspective: A Discussion of Common Misconceptions. *Perspectives on Psychological Science* 14:6. DOI: 10.1177/1745691619861372.

Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie i Yorick Wilks (2006). A Closer Look at Skip-gram Modelling. Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk i Daniel Tapias, ur. *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC*. Genoa, Italy: European Language Resources Association, 1222-1225.

Harnad, Stevan (1990). The Symbol Grounding Problem. *Physica D42*: 335-346.

Harris, Zellig (1968). *Mathematical Structures of Language*. New York: Interscience Publishers.

Herağdelen, Amaç i Marco Marelli (2017). Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive science* 41(4): 976-995.

Hintzman, Douglas L. (1986). "Schema Abstraction" in a Multiple-Trace Memory Model. *Psychological Review* 94(4): 411-428.

Holderbaum, Candice Steffen i Jerusa Fumagalli de Salles (2011). Semantic Priming Effects in a Lexical Decision Task: Comparing third Graders and College Students in two Different Stimulus Onset Asynchronies. *The Spanish Journal of Psychology* 14(2): 589-599.

Horvatić, Krešimir (2004). *Linearna algebra*. Zagreb: Golden Marketing – Tehnička knjiga.

Hutchinson, Sterling i Max M. Louwerse (2013). What's UP can be Explained by Language Statistics. Markus Knauff, Miahel Pauen, Natalie Sebanz, Ipke Wachsmuth, ur. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*. Berlin: Cognitive Science Society, 2596-2601.

Hutchison, Keith A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review* 10(4): 785-813.

Hutchison, Keith A., David A. Balota, James J. Neely, Michael J. Cortese, Emily R. Cohen-Shikora, Chi-Shing Tse, Melvin J. Yap, Jesse J. Bergson, Dale Niemeier i Erin Buchanan (2013). The semantic priming project. *Behaviour Research Methods* 45(4): 1099-114.

Hutchison, Keith A., David A. Balota, Michael J. Cortese i Jason M. Watson (2008). Predicting semantic priming at the item level. *Quarterly journal of experimental psycholog.* 61(7): 1036-66.

Huynh, Juliet i Naoko Witzel (2018). Associative networks from L2 words in early and late Vietnamese-English bilinguals. *Journal of Second Language Studies* 1(1): 199-230.

Illes, Judy, Wendy S. Francis, John E. Desmond, John D. E. Gabrieli, Gary H. Glover, Russel Poldrack, Christine J. Lee i Anthony D. Wagner (1999). Convergent Cortical Representation of Semantic Processing in Bilinguals. *Brain and Language* 70: 347-363.

Izura, Cristina i Andrew W. Ellis (2002). Age of acquisition in word recognition and production in first and second languages. *Psicologica* 23: 245-281.

Izura, Cristina, Miguel A. Pérez, Elizabeth Agallou, Victoria C. Wright, Javier Marín, Hans Stadtgaheb-González i Andrew E. Ellis (2011). Age/order of acquisition effects and the cumulative learning of foreign words: A word training study. *Journal of memory and language* 64: 35-58.

Jeong, Hyeonjeong, Motoaki Sugiura, Yuko Sassa, Keisuke Wakusawa, Kaoru Horie, Shigeru Sato i Rjuta Kawashima (2010). Learning second language vocabulary: Neural dissociation of situation-based learning and text-based learning. *NeuroImage* 50(2): 802-809.

Johns, Brendan T. i Michael N. Jones (2015). Generating structure from experience: A retrieval-based model of language processing. *Canadian journal of experimental psychology/Revue canadienne de psychologie experimentale* 69(3): 233-251.

Jones, Michael N., i Douglas J. K. Mewhort (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review* 114: 1-37.

Jones, Michael N., Walter Kintsch i Douglas J. K. Mewhort (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language* 55: 534-552.

Jurafsky, Dan, James H. Martin (2020). *Speech and Language Processing*.

Keuleers, Emmanuel i David A. Balota (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *Quarterly journal of experimental psychology* 68(8): 1457-1468.

Keuleers, Emmanuel i Marc Brysbaert (2010). Wuggy. A multilingual pseudoword generator. *Behaviour Research Method* 42: 627-633.

Keuleers, Emmanuel, Michael Stevens, Paweł Mandera i Marc Brysbaert (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly journal of experimental psychology* 68(8): 1665-1692.

Kiela, Douwe i Stephen Clark (2015). Multi- and Cross-Modal Semantics Beyond Vision: Grounding in Auditory Perception. Lluís Marquez, Chris Callison-Burch i Jian Su, ur. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2461–2470.

Kiela, Douwe, Felix Hill, Anna Korhonen i Stephen Clark (2014). Improving Multi-Modal Representations Using Image Dispersion: Why Less is Sometimes More. Cristian Danescu-Miculescu-Mizil, Jacob Eisenstein, Kathleen McKeown i Noah A. Smith, ur. *Proceedings of ACL*. Baltimore: Association for Computational Linguistics, 835-841.

Kiela, Douwe, Luana Bulat i Stephen Clark (2015). Grounding Semantics in Olfactory Perception. Chengqing Zong i Michael Strube, ur. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*. Beijing: Association for Computational Linguistics, 231–236.

Kiela, Douwe (2017). *Deep embodiment: grounding semantics in perceptual modalities*. Doktorska disertacija.

Kintsch, Walter (2008). Symbol systems and perceptual representation. Manuel de Vega, Arthur Glenberg i Arthur Graeser, ur. *Symbols and Embodiment*. Oxford: Oxford University Press, 145-164.

Koriat, Asher (1981). Semantic facilitation in lexical decision as a function of prime-target association. *Memory and Cognition* 9: 587-598.

Kovačec, August (2001). Ferdinand de Saussure i strukturalizam. Zrinjka Glovacki-Bernardi, ur. *Uvod u lingvistiku*. Zagreb: Školska knjiga, 75-154.

Kroll, Judith F. i Alexandra Scholl (1992). Lexical and Conceptual Memory in Fluent and Nonfluent Bilinguals. *Advances in Psychology* 83: 191-204.

Kroll, Judith F. i Erika Stewart (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language* 33: 149–174.

Kroll, Judith F. i Natasha Tokowicz (2005). Models of Bilingual Representation and Processing: Looking Back and to the Future. Judith F. Kroll i Annette M. B. de Groot, ur.

Handbook of bilingualism: Psycholinguistic approaches. New York: Oxford University Press, 532-553.

Kroll, Judith F., Janet G. van Hell, Natasha Tokowicz i David W. Green (2010). The Revised Hierarchical Model: A critical review and assessment. *Bilingualism: Language and Cognition* 13(3): 373-381.

Kučera, Henry i W. Nelson Francis (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Kumar, Abhilasha A. (2020). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*. DOI: 13.3758/s13423-020-01792-x.

Lake, Brenden M. i Gregory L. Murphy (2020). Word meaning in minds and machines. arXiv:2008.01766v2 [cs.CL].

Lake, Brenden M., Ruslan Salakhutdinov i Joshua B. Tenenbaum (2015). Human-level concept learning through probabilistic program induction. *Science* 350: 1332-1338.

Lakoff, George i Mark Johnson (1999). *Philosophy in the Flesh: the Embodied Mind and its Challenge to Western Thought*. New York: Basic Books.

Landau, Barbara, Linda B. Smith i Susan S. Jones (1988). The importance of shape in early lexical learning. *Cognitive Development* 3(3): 299-321.

Landauer, Thomas K. (1999). Latent Semantic Analysis (LSA), a disembodied learning machine, acquires human word meaning vicariously from language alone. Lawrence W. Barsalou. Perceptual symbol systems. *Behavioral and Brain Sciences* 22(4): 624-625.

Landauer, Thomas K. (2007). LSA as a theory of meaning. Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, Walter Kintsch, ur. *Handbook of latent semantic analysis*. Mahwah, N.J.: Lawrence Erlbaum Association, 3-34.

Landauer, Thomas K. i Susan T. Dumais (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review* 104(2): 211-240.

Landauer, Thomas K., Peter W. Foltz i Darrell Laham (1998). An introduction to latent semantic analysis. *Discourse Processes* 25(2-3): 259-284.

Langacker, Ronald W. (2008). *Cognitive Grammar. A Basic Introduction*. New York: Oxford University Press.

Lange, Kristian, Simone Kühn i Elisa Filevich (2015). "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PLOS ONE* 10.7. e0130834.

Lee, Yoonhyoung, Euna Jang i Wonil Choi (2018). L2-L1 Translation Priming Effects in a Lexical Decision Task: Evidence From Low-Proficient Korean-English Bilinguals. *Frontiers in Psychology*. doi: 10.3389/fpsyg.2018.002967.

Lemhöfer Kristin i Ton Dijkstra (2004). Recognizing cognates and interlingual homographs: Effects of cognate similarity in language-specific and generalized lexical decision. *Memory & Cognition* 32(4): 533-550.

Lenci, Alessandro (2008). Distributional semantics in linguistic and cognitive research. *The Italian Journal of Linguistics* 20: 1-32.

Lenz, Siegfried (1999). *Sat njemačkoga*. Zagreb: Školska knjiga.

Leong, Chee Wee i Rada Mihalcea (2011). Going Beyond Text: A Hybrid Image-Text Approach for Measuring Word Relatedness. Haifeng Wang i David Yarowsky, ur. *Proceedings of the 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, 1403-1407.

Letsche, Todd A. i Michael W. Berry (1997). Large-Scale Information Retrieval with Latent Semantic Indexing. *Information Sciences* 100(1-4): 105-137.

Levin, Beth (1993). *English Verb Classes and Alternations. A Preliminary Investigation*. Chicago: University of Chicago Press.

Levy, Omer, Yoav Goldberg i Ido Dagan (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Computational Linguistics* 3: 211-225.

Levy, Omer i Yoav Goldberg (2014). Neural Word Embedding as Implicit Matrix Factorization. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence i K. Q. Weinberger, ur. *Proceedings of the 27th International Conference on Neural Information Processing Systems*. Cambridge: MIT Press, 2177-2185.

Lévy, Nicolas i François Grosjean (2008). The Lévy and Grosjean BIMOLA model. François Grosjean, ur. *Studying Bilinguals*. Oxford: University Press, 201-212.

Li, Ping i Igor Farkaš (2002). A self-organizing connectionist model of bilingual processing. R. Heredia i J. Altarriba, ur. *Bilingual Sentence Processing*. North Holland: Elsevier Science Publisher, 59-85.

Li, Ping, Curt Burgess i Kevin Lund (2000). The Acquisition of Word Meaning through Global Lexical Co-occurrences. Eve V. Clark, ur. *Proceedings of the Thirtieth Stanford Child Language Research Forum*. Stanford: Centre for the Study of Language & Information, 167-178.

Lin, Dekang i Patrick Pantel (2001). DIRT – discovery of inference rules from text. Foster Provost i Ramakrishnan Srikant, ur. *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001*. New York: Association for Computing Machinery, 323-328.

Lin, Dekang (1998). Automatic Retrieval and Clustering of Similar Words. Christian Boitet i Pete Whitelock, ur. *ACL '98 Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. 2*. Montreal: Association for Computational Linguistics, 768-774.

Linck, Jared A., Judith F. Kroll i Gretchen Sunderman (2009). Losing Access to the Native Language While Immersed in a Second Language: Evidence for the Role of Inhibition in Second-Language Learning. *Psychological Science* 20(12): 1507-1515.

Lindemann, Oliver, Ahmad Alipour i Martin H. Fischer (2011). Finger Counting Habits in Middle Eastern and Western Individuals: An Online Survey. *Journal of Cross-Cultural Psychology* 42(4): 566-578.

Loetscher, Tobias, Urs Schwarz, Michele Schubiger i Peter Brugger (2008). Head turns bias brain's internal random generator. *Current Biology* 18(2): R60-R62.

Lotto, Lorella i Annette M. B. de Groot (1998). Effects of Learning Method and Word Type on Acquiring in an Unfamiliar Language. *Language Learning* 48(1): 31-69.

Louwerse, Max i Patrick Jeuniaux (2008). Language comprehension is both embodied and symbolic. Manuel de Vega, Arthur Glenberg i Arthur Graesser, ur. *Symbols and Embodiment*. Oxford: Oxford University Press, 309-325.

Louwerse, Max M. (2011). Symbol Interdependency in Symbolic and Embodied Cognition. *Topics in Cognitive Sciences* 3(2): 273-302.

Louwerse, Max M. i Louise Connell (2011). A Taste of Words: Linguistic Context and Perceptual Simulation Predict the Modality of Words. *Cognitive Science* 35(2): 381-398.

Louwerse, Max M. i Rolf A. Zwaan (2009). Language Encodes Geographical Information. *Cognitive Science* 35: 381-398.

Lucas, Margery (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review* 7(4): 618-630.

Lund, Kevin i Curt Burgess (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour Research Methods, Instruments, & Computers* 28(2): 203-208.

Lund, Kevin, Curt Burgess i Ruth Ann Atchley (1995). Semantic and Associative Priming in High-Dimensional Semantic Space. Johanna D. Moore i Jill Fain Lehman, ur. *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*. Pittsburgh: University of Pittsburgh, 660-665.

Mandera, Paweł, Emmanuel Keuleers i Marc Brysbaert (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language* 92: 57-78.

Mandera, Paweł, Emmanuel Keuleers, Zofia Wodniecka i Marc Brysbaert (2015). Sublex-pl: subtitlebased word frequency estimates for Polish. *Behavior research methods* 47(2): 471-483.

Mao, Jiayuna, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum i Jiajun Wu (2019). The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. arXiv:1904.12584 [cs.CV].

Marian, Viorica, Henrike K. Blumenfeld i Margarita Kaushanskaya (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q). Assessing Language Profiles in Bilinguals and Multilinguals. *Journal of Speech, Language, and Hearing Research* 50: 940-967.

Marian, Viorica, Michael Spivey i Joy Hirsch (2003). Shared and separate systems in bilingual language processing: Converging evidence from eye-tracking and brain imaging. *Brain and Language* 86: 70-82.

Martin, Dian I. i Michael W. Berry (2007). Mathematical Foundations Behind Latent Semantic Analysis. Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, Walter Kintsch, ur. *Handbook of Latent Semantic Analysis*. NY & London: Routledge.

Matheson, Heath E. i Lawrence W. Barsalou (2018). Embodiment and Grounding in Cognitive Neuroscience. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* 3: 1-27.

Mathôt, Sebastiaan, Daniel Schreij i Jan Theeuwes (2012). OpenSesame: An open-source, graphical experiment builder for social sciences. *Behavior Research Methods* 44: 314-324.

McClelland, James L. i Jeffrey L. Elman (1986). The TRACE model of speech perception. *Cognitive Psychology* 18(1): 1-86.

Meade, Gabriela, Katherine J. Midgley i Phillip J. Holcomb (2018). An ERP Investigation of L2-L1 Translation Priming in Adult Learners. *Frontiers in Psychology*. doi: 10.3389/fpsyg.2018.00986

Mihaljević Djigunović, Jelena i Renata Geld (2002). English in Croatia Today: Opportunities for Incidental Vocabulary Acquisition. *Studia Romanica et Anglica Zagabiensia: Revue publiée par les Sections romane, italienne et anglaise de la Faculté des Lettres de l'Université de Zagreb* 47-48: 335-352.

Mikolov, Tomas, Kai Chen, Greg Corrado i Jeffrey Dean (2013a). Efficient Estimation of Word Representations in Vector Space. *Computation and Language (cs.CL)*. arXiv:1301.3781 [cs.CL].

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado i Jeffrey Dean (2013b). Distributed Representations of Words and Phrases and their Compositionality. *Computation and Language (cs.CL)*. arXiv:1310.4546 [cs.CL].

Mikolov, Tomas; Wen-Tau Yih i Geoffrey Zweig (2013). Linguistic Regularities in Continuous Space Word Representations. Lucy Vanderwende, Hal Daume III i Katrin Kirchhoff, ur. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta: Association for Computational Linguistics, 746-751.

Miller, George A. i Jennifer A. Selfridge (1949). Verbal context and the recall of meaningful material. *American Journal of Psychology* 63: 176-185.

Miller, George A. i Walter G. Charles (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6: 1-28.

Miller, Ralph R., Robert C. Barnet i Nicholas J. Grahame (1995). Assessment of the Rescorla-Wagner Model. *Psychological Bulletin* 117(3): 363-386.

Moore, Johanna C. i John R. Surber (1992). Effects of Context and Keyword Methods on Second Language Vocabulary Acquisition. *Contemporary Educational Psychology* 17(3): 286-292.

Morrison, Catriona M. i Andrew W. Ellis (2000). Real age of acquisition effect in word naming and lexical decision. *British Journal of Psychology* 91: 167-180.

Moseley Rachel, Markus Kiefer i Friedemann Pulvermüller (2016). Grounding and embodiment of concepts and meaning: A neurobiological perspective. Yann Coello i Martin H. Fischer, ur. *Perceptual and Emotional Embodiment: Foundations of Embodied Cognition*. New York: Routledge, 93-113.

Muñoz, Carmen (2008). Age-related differences and second language learning practice. Robert DeKeyser, ur. *Practice in a Second Language: Perspectives from Applied Linguistics and Cognitive Psychology*. New York: Cambridge University Press, 229-255.

Murdock, Bennet B. (1982). A Theory for the Storage and Retrieval of Item and Associative Information. *Psychological Review* 89(6): 609-626.

Nadalini, Andrea, Marco Marelli, Roberto Bottini i Davide Crepaldi (2018). Local associations and semantic ties in overt and masked semantic priming. Elena Cabrio, Alessandro Mazzei i Fabio Tamburini, ur. *Proceedings of the fifth Italian conference on computational linguistics. CLIC-IT*. Torino: Accademia University Press.

Nakayama, Mariko, Stephen J. Lupker i Yoshihiro Itaguchi (2018). An examination of L2-L1 noncognate translation priming in the lexical decision task: Insights from distributional and frequency-based analyses. *Bilingualism: Language and Cognition* 21(2): 265-277.

Navracsis, Judit (2000). Word association test among bilinguals. *Strani jezici* 29(3): 143-150.

Neely, James H., Dennis E. Keefe i Kent L. Ross (1989). Semantic priming in the lexical decision task: Roles of prospective prime-generated expectancies and retrospective semantic matching. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15(6): 1003-1019.

New, Boris, Marc Brysbaert, Jean Veronis i Christophe Pallier (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics* 28: 661-667.

Nida, Eugene, A. (1975). *A Componential Analysis of Meaning: An Introduction to Semantic Structures*. The Hague: Mouton.

Onifer, William i David A. Swinney (1981). Access lexical ambiguities during sentence comprehension: Effects of frequency of meaning and contextual bias. *Memory and Cognition* 9(3): 225-236.

Osgood, Charles E., George J. Suci i Percy H. Tannenbaum (1957). *The Measurement of Meaning*. Urbana: University of Illinois Press.

Ossorio, Peter G. (1966). A multivariate procedure for automatic document indexing and retrieval. *Multivariate Behavioral Research*: 479-524.

Paivio, Allan (1986). *Mental representations: a dual coding approach*. Oxford, England: Oxford University Press.

Paperno, Denis; Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda i Raquel Fernández (2016). The LAMBADA dataset. Word Prediction requiring a broad discourse context. arXiv:1606.06031 [cs.CL].

Paradis, Michel (1997). The cognitive neuropsychology of bilingualism. Annette M. B. de Groot i Judith F. Kroll, ur. *Tutorials in Bilingualism. Psycholinguistic Perspectives*. New York: Lawrence Erlbaum Associates, 311-354.

Peal, Elizabeth i Wallace E. Lambert (1962). The relation of bilingualism to intelligence. *Psychological Monographs: General and Applied* 76(27): 1-23.

Peirce, Jonathan, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman i Jonas Kristoffer Lindeløv (2019). PsychoPy2. Experiments in behavior made easy. *Behavior Research Methods* 51: 195-203.

Perani, Daniela, Eraldo Paulesu, Nuria Sebastian Galles, Emmanuel Dupoux, Stanislas Dehaene, Valentino Bettinardi, Stefano F. Cappa, Ferruccio Fazio i Jacques Mehler (1998). The Bilingual Brain: Proficiency and age of acquisition of the second language. *Brain* 121: 1841-1852.

Perea, Manuel, Jon Andoni Duñabeitia i Manuel Carreiras (2008). Masked associative/semantic priming effects across languages with highly proficient bilinguals. *Journal of Memory and Language* 58: 916-930.

Pereira, Francisco, Samuel Gershman, Samuel Ritter i Matthew Botvinick (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology* 33(3): 175-190.

Perfetti, Charles A. (1998). The Limits of Co-Occurrence: Tools and Theories in Language Research. *Discourse Processes* 25(2-3): 363-377.

Peters, Elke i Stuart Webb (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*: 1-27.

Pezzulo, Giovanni, Lawrence W. Barsalou, Angelo Cangelosi, Martin H. Fischer, Ken McRae i Michael J. Spivey (2012). Computational Grounded Cognition: a new alliance between grounded cognition and computational modeling. *Frontiers in Psychology* 3. Čl. 612.

Plate, Tony (1995). Holographic Reduced Representations. *IEEE Transactions on Neural Networks* 6(3): 623-641.

Plate, Tony (2003). *Holographic Reduced Representations*. Stanford: CSLI Publications.

Potter, Mary C., Kwok-Fai So, Barbara von Eckardt i Laurie B. Feldman (1984). Lexical and conceptual representation in beginning and proficient bilinguals. *Journal of Verbal Learning and Verbal Behavior* 23(1): 23-38.

Pulvermüller, Friedemann (2005). Brain Mechanisms Linking Language and Action. *Nature reviews Neuroscience* 6(7): 576-582.

Pulvermüller, Friedemann (2013). Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits. *Brain and Language* 127(1): 86-103.

Pustejovsky, James (2001). Type construction and the logic of concepts. Bouillon Pierette i Federica Busa, ur. *The Syntax of Word Meaning*. Cambridge: Cambridge University Press, 91-123.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. Austria. URL <https://www.R-project.org/>.

Racine, John P., George Higginbotham i Ian Munby (2014). Exploring non-native norms: A new direction in word association research. *Vocabulary Education and Research Bulletin* 3(2): 13-15

Recchia, Gabriel i Michael N. Jones (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods* 41(3): 647-656.

Rehder, Bob, M. E. Schreiner, Michael B. W. Wolfe, Darrell Laham, Thomas K. Landauer i Walter Kintsch (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes* 25(2-3): 337-354.

Rescorla, Robert A. i Allan R Wagner (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. Abraham H. Black i William F. Prokasy, ur. *Classical Conditioning II: Current Theory and Research*. New York: Appleton-Century-Crofts, 64-99.

Riordan, Brian i Michael N. Jones (2011). Redundancy in perceptual and linguistic experience: comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science* 3(2): 303-345

Rumshisky, Anna, Marc Verhagen i Jessica L. Moszkowicz (2009). The Holy Grail of Sense Definition. Creating a Sense-Disambiguated Corpus from Scratch. Pisa.

Sahlgren, Magnus (2006). The Word-Space Model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Doktorska disertacija.

Sahlgren, Magnus (2008). The Distributional Hypothesis. From context to meaning: Distributional models of the lexicon in linguistics and cognitive science. *Special issue of the Italian Journal of Linguistics, Rivista di Linguistica* 20(1): 33-53.

Sahlin, Brooke H., Matthew G. Harding i John G. Seamon (2005). When do false memories cross language boundaries in English-Spanish bilinguals? *Memory & Cognition* 33(8): 1414-1421.

Salton, Gerard (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. River: Prentice Hall.

Samir, Ferhan, Barend Beekhuizen i Suzanne Stevenson (2020). Untangling Semantic Similarity: Modeling Lexical Processing Experiments with Distributional Semantic Models. Creative Commons Attribution 4.0 International License.

Schepens, Job. J., Frans van der Silk i Roeland van Hout (2015). L1 and L2 Distance Effects in Learning L3 Dutch. *Language Learning* 66(1): 224-256.

Scholz, Barbara C., Francis Jeffrey Pelletier i Geoffrey K. Pullum (2020). Philosophy of Language. Edward N. Zalta, ur. *The Stanford Encyclopedia of Philosophy*. Summer 2020 Edition. <https://plato.stanford.edu/archives/sum2020/entries/linguistics/> [24.1.2021.]

Schoonbaert, Sofie, Wouter Duyck, Marc Brysbaert i Robert J. Hartsuiker (2009). Semantic and translation priming from a first language to a second and back: Making sense of the findings. *Memory & Cognition* 37: 569-586.

Searl, John (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences* 3(3): 417-457.

Serafin, Riccardo i Di Eugenio, Barbara (2004). FLSA: Extending latent semantic analysis with features for dialogue act classification. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. 692-699.

Shapiro, Lawrence (2008). Symbolism, Embodied Cognition, and the Broader Debate. Manuel de Vega, Arthur Glenberg i Arthur Graeser, ur. *Symbols and Embodiment*. Oxford: Oxford University Press, 57-74.

Silverberg, Stu i Arthur G. Samuel (2004). The effect of age of second language acquisition on the representation and processing of second language words. *Journal of Memory and Language* 51: 381-398.

Snodgrass, Joan Gay i Mary Vanderwart (1980). A Standardized Set of 260 Pictures. Norms for Name Agreement, Image Agreement, Familiarity, and Visual Complexity. *Journal of Empirical Psychology. Human Learning and Memory* 6(2): 174-215.

Socher, Richard, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng i Christopher D. Manning (2011). Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. NIPS'11 Proceedings of the 24th International Conference on Neural Information Processing Systems. Granada: Curran Associates, 801-809.

Strang, Gilbert (2009). *Introduction to Linear Algebra*. Wellesley: Cambridge Press.

Sunderman, Gretchen L. (2011). Conceptual mediatoin in second language learner. Pavel Trofimovich i Kim McDonough, ur. *Applying priming methods to L2 learning, teaching and reseach: Insights from Psycholinguistics*. Amsterdam: Benjamins, 219-237.

Thomas, Michael S. C. (1997). Distributed representations and the bilingual lexicon: One store or two? J. Bullinaria, D. Glasspool i G. Houghton, ur. *Proceedings of the Fourth Annual Neural Computation and Psychology Workshop*. London: Springer. 240-253.

Thomas, Michael S. C. i Walter J. B. van Heuven (2005). Computational Models of Bilingual Comprehension. Judith F. Kroll i Annette M. B. de Groot, ur. *Handbook of bilingualism: Psycholinguistic approaches*. New York: Oxford University Press, 202-225.

Tillman, Richard, Vivek Datla, Sterling Hutchinson i Max M. Louwerse (2012). From Head to Toe: Embodiment Through Statistical Linguistic Frequencies. Naomi Miyake, David Peebles i Richard P. Cooper, ur. *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*. Austin: Cognitive Science Society, 2434-2439.

Titone, Debra, Maya Libben, Julie Mercier, Veronica Whitford i Irina Pivneva (2011). Bilingual Lexical Access During L1 Sentence Reading: The Effects of L2 Knowledge, Semantic Constraint, and L1-L2 Intermixing. *Journal of Experimental Psychology: Learning Memory, and Cognition* 37(6): 1412-1431.

Trofimovich, Pavel (2011). When three equals tree: Examining the nature of phonological entries in L2 lexicons of Quebec speakers of English. Pavel Trofimovich i Kim McDonough, ur. *Applying priming methods to L2 learning, teaching and reseach: Insights from Psycholinguistics*. Amsterdam: Benjamins, 105-129.

Turney, Peter D i, Patrick Pantel (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37: 141-188.

Ulrich, Martin, Klaus Hoenig, Georg Grön i Markus Kiefer (2013). Brain Activation during Masked and Unmasked Semantic Priming: Commonalities and Differences. *Journal of Cognitive Neuroscience* 25(12): 2216-2229.

Van Hell, Janet G. i Andrea Candia Mahn (1997). Keyword Mnemonics Versus Rote Rehearsal: Learning Concrete and Abstract Foreign Words by Experienced and Inexperienced Learners. *Language Learning* 47(3): 507-546.

Van Hell, Janet G. i Annette M. B. de Groot (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition*. 1(3): 193-211.

Van Hell, Janet G. i Darren Tanner (2012). Second Language Proficiency and Cross-Language Lexical Activation. *Language Learning* 62: 148-171.

Van Hell, Janet G. i Ton Dijkstra (2002). Foreign language knowledge can influence native language performance in exclusively native contexts. *Psychonomic Bulletin & Review* 9(4): 780-789.

Van Heuven, Walter J. B., Ton Dijkstra i Jonathan Grainger (1998). Orthographic Neighborhood Effects in Bilingual Word Recognition. *Journal of Memory and Language* 39(3): 458-483.

Van Heuven, Walter J.B., Herbert Schriefers, Ton Dijkstra i Peter Hagoort (2008). Language Conflict in the Bilingual Brain. *Cereb Cortex* 18(11): 2706-2716.

Vezmar, Nika (2017). Norme slobodnih asocijacija za uzorak riječi na hrvatskom jeziku. Diplomski rad. Filozofski fakultet. Zagreb.

Vigliocco, Gabriella, Stavroula-Thaleia Kousta, Pasquale Anthony Della Rosa, David P. Vinson, Marco Tettamanti, Joseph T. Devlin i Stefano F. Cappa (2013). The Neural Representation of Abstract Words: The Role of Emotions. *Cerebral Cortex* 24(7): 1767-1777.

Vijeće Europske unije (2001). *Common European Framework of Reference for Languages: Learning teaching, assessment*. Cambridge: Cambridge University Press.

Von Studnitz, Roswitha E. i David W. Green (2002=). Interlingual homograph interference in German-English bilinguals: Its modulation and locus of control. *Bilingualism: Language and Cognition* 5(1): 1-23.

Vulić, Ivan i Marie-Francine Moens (2016). Bilingual Distributed Word Representations from Document-Aligned Comparabel Data. *Journal of Artificial Intelligence Research* 55: 953-994.

Wang, Xin i Kenneth Forster (2010). Masked translation priming with semantic categorization: Testing the Sense Model. *Bilingualism: Language and Cognition*: 1-14.

Widdows, Dominic (2004). *Geometry and meaning*. Stanford: Centre for Study of Language and Information.

Wiemer-Hastings, Peter i Iraide Zipitria (2001). Rules for Syntax, Vectors for Semantics. Lila Gleitman i Aravind K. Joshi, ur. *Proceeding of the Twenty-Second Annual Conference of the Cognitive Science Society*. Pennsylvania: Psychology Press, 989-1015.

Williams, John N. i Amanda Cheung (2011). Using priming to explore early word learning. Pavel Trofimovich i Kim McDonough, ur. *Applying priming methods to L2 learning, teaching and reseach: Insights from Psycholinguistics*. Amsterdam: Benjamins. 73-103.

Wilson, Andrew D. i Sabrina Golonka. (2013). Embodied Cognition is Not What you Think it is. *Frontiers in Psychology* 4(58). <https://doi.org/10.3389/fpsyg.2013.00058>

Wolfe, Michael B. W., M. E. Schreiner, Bob Rehder i Darrell Laham (1998). Learning From Text: Matching Readers and Text by Latent Semantic Analysis. *Discourse Processes* 25(2-3): 309-336.

Xia, Violet i Sally Andrews (2015). Masked translation priming asymentry in Chinese-English bilinguals: Making sense of the Sense Model. *The Quaterly Journal of Experimental Psychology* 68(2): 294-325.

Kada biste morali razgovarati s nekom osobom na stranom jeziku, u kojem biste postotku odabrali svaki pojedini strani jezik? (Zbroj bi trebao iznositi 100%.)

1. strani jezik _____
 2. strani jezik _____
 3. strani jezik _____

Kako procjenjujete svoje poznavanje stranoga jezika?

1. strani jezik

Čitanje

Izvršno	Vrlo dobro	Dobro	Dovoljno	Nedovoljno
---------	------------	-------	----------	------------

Govorenje

Izvršno	Vrlo dobro	Dobro	Dovoljno	Nedovoljno
---------	------------	-------	----------	------------

Razumijevanje

Izvršno	Vrlo dobro	Dobro	Dovoljno	Nedovoljno
---------	------------	-------	----------	------------

2. strani jezik

Čitanje

Izvršno	Vrlo dobro	Dobro	Dovoljno	Nedovoljno
---------	------------	-------	----------	------------

Govorenje

Izvršno	Vrlo dobro	Dobro	Dovoljno	Nedovoljno
---------	------------	-------	----------	------------

Razumijevanje

Izvršno	Vrlo dobro	Dobro	Dovoljno	Nedovoljno
---------	------------	-------	----------	------------

3. strani jezik

Čitanje

Izvršno	Vrlo dobro	Dobro	Dovoljno	Nedovoljno
---------	------------	-------	----------	------------

Govorenje

Izvršno	Vrlo dobro	Dobro	Dovoljno	Nedovoljno
---------	------------	-------	----------	------------

Razumijevanje

Izvršno	Vrlo dobro	Dobro	Dovoljno	Nedovoljno
---------	------------	-------	----------	------------

Koliko su sljedeći čimbenici doprinijeli Vašem poznavanju njemačkoga jezika?

Interakcija s prijateljima

Uopće nije doprinijelo	Uglavnom nije doprinijelo	Niti je doprinijelo niti nije doprinijelo	Uglavnom je doprinijelo	Izrazito je doprinijelo
------------------------	---------------------------	---	-------------------------	-------------------------

Interakcija s obitelji

Uopće nije doprinijelo	Uglavnom nije doprinijelo	Niti je doprinijelo niti nije doprinijelo	Uglavnom je doprinijelo	Izrazito je doprinijelo
------------------------	---------------------------	---	-------------------------	-------------------------

Gledanje televizije, filmova ili internetskih sadržaja

Uopće nije doprinijelo	Uglavnom nije doprinijelo	Niti je doprinijelo niti nije doprinijelo	Uglavnom je doprinijelo	Izrazito je doprinijelo
------------------------	---------------------------	---	-------------------------	-------------------------

Čitanje

Uopće nije	Uglavnom nije	Niti je doprinijelo	Uglavnom je	Izrazito je
------------	---------------	---------------------	-------------	-------------

doprinijelo	doprinijelo	niti nije doprinijelo	doprinijelo	doprinijelo
Slušanje radija i glazbe				
Uopće nije doprinijelo	Uglavnom nije doprinijelo	Niti je doprinijelo niti nije doprinijelo	Uglavnom je doprinijelo	Izrazito je doprinijelo
Samostalno učenje				
Uopće nije doprinijelo	Uglavnom nije doprinijelo	Niti je doprinijelo niti nije doprinijelo	Uglavnom je doprinijelo	Izrazito je doprinijelo

Koliko ste trenutačno izloženi svakom od sljedećih čimbenika na njemačkom jeziku?

Interakcija s prijateljima

Uopće nisam izložen/a	Uglavnom nisam izložen/a	Niti jesam izložen/a niti nisam izložen/a	Uglavnom sam izložen/a	Izrazito sam izložen/a
-----------------------	--------------------------	---	------------------------	------------------------

Interakcija s obitelji

Uopće nisam izložen/a	Uglavnom nisam izložen/a	Niti jesam izložen/a niti nisam izložen/a	Uglavnom sam izložen/a	Izrazito sam izložen/a
-----------------------	--------------------------	---	------------------------	------------------------

Gledanje televizije, filmova ili internetskih sadržaja

Uopće nisam izložen/a	Uglavnom nisam izložen/a	Niti jesam izložen/a niti nisam izložen/a	Uglavnom sam izložen/a	Izrazito sam izložen/a
-----------------------	--------------------------	---	------------------------	------------------------

Čitanje

Uopće nisam izložen/a	Uglavnom nisam izložen/a	Niti jesam izložen/a niti nisam izložen/a	Uglavnom sam izložen/a	Izrazito sam izložen/a
-----------------------	--------------------------	---	------------------------	------------------------

Slušanje radija i glazbe

Uopće nisam izložen/a	Uglavnom nisam izložen/a	Niti jesam izložen/a niti nisam izložen/a	Uglavnom sam izložen/a	Izrazito sam izložen/a
-----------------------	--------------------------	---	------------------------	------------------------

Samostalno učenje

Uopće nisam izložen/a	Uglavnom nisam izložen/a	Niti jesam izložen/a niti nisam izložen/a	Uglavnom sam izložen/a	Izrazito sam izložen/a
-----------------------	--------------------------	---	------------------------	------------------------

Koliko ste godina učili njemački jezik u školi? _____

Jeste li ikada učili njemački jezik u školi stranih jezika? Da Ne

Ako jeste učili njemački u školi stranih jezika, koliko ste ga godina učili? (Zaokružite na punu godinu.) _____

Jeste li ikada boravili dulje od pola godine bez prestanka u zemljama njemačkoga govornog područja? Da Ne


Koliko ste godina boravili ondje? _____

Procijenite koliko se sati tjedno koristite njemačkim jezikom (slušanje, govorenje, čitanje, pisanje)? _____

Procijenite koliko vremena tjedno provodite "na" njemačkom, a koliko "na" hrvatskom jeziku? Koliko, eventualno, "na" nekom drugom jeziku?

Njemački jezik (u %) _____
Hrvatski jezik (u %) _____
Drugi jezik (u %) _____
Koji je to drugi jezik? _____

16.2. Sličice koje su ispitanici vidjeli tijekom eksperimenta

<p>Dobrodošli u eksperiment!</p> <p>Eksperiment se sastoji od dva dijela: njemački dio i hrvatski dio. Eksperiment će se provesti na sljedeći način. Na sredini će se ekrana pokazati točka koja služi kao znak fiksacije.</p> <p>Za nastavak pritisni tipku razmak.</p>	
<p>Potom će na mjestu znaka fiksacije (točka) na ekranu biti vrlo kratko predstavljena postojeća riječ u hrvatskom jeziku, zatim će biti predstavljen određen niz slova. Neki nizovi slova predstavljat će postojeće riječi (imenice) njemačkog jezika, neki će predstavljati nepostojeće riječi njemačkog jezika. Slijede primjeri za nizove slova koje predstavljaju postojeće i nepostojeće riječi u njemačkom jeziku.</p> <p>Za nastavak pritisni tipku razmak.</p>	<p>Fernseher</p>
<p>Ili ako se radi o nepostojećoj riječi u njemačkom jeziku:</p>	<p>Kascheltust</p>
<p>Vaš je zadatak ŠTO JE BRŽE MOGUĆE odlučiti je li ponuđeni niz slova riječ u NJEMAČKOM jeziku ili nije.</p> <p>Mjeri se Vaša reakcija na drugu, njemačku riječ.</p> <p>Pritisnite tipku j, ako ponuđeni niz slova JEST postojeća riječ u njemačkom jeziku. Pritisnite tipku f, ako ponuđeni niz slova</p>	<p>Molim Vas da postavite desni kažiprst na tipku j, a lijevi kažiprst na tipku f.</p> <p>Čim pritisnete tipku j ili f, pojavit će se sljedeći niz slova.</p> <p>Najprije ćete napraviti nekoliko zadataka za vježbu.</p> <p>Molim Vas da budete što brži i što točniji.</p> <p>Kad budete spremni, pritisnite tipku j ili f i</p>

<p>NIJE postojeća riječ u njemačkom jeziku.</p> <p>Za nastavak pritisni tipku razmak.</p>	<p>vježba može krenuti.</p>
<p>Ovo je bila vježba.</p> <p>U nastavku ćemo prijeći na eksperiment.</p> <p>Trebat ćete učiniti isto kao i do sada.</p> <p>Za nastavak pritisni tipku razmak.</p>	<p>Vaš je zadatak ŠTO JE BRŽE MOGUĆE odlučiti je li ponudeni niz slova riječ u NJEMAČKOM jeziku ili nije.</p> <p>Mjeri se Vaša reakcija na drugu, njemačku riječ.</p> <p>Pritisnite tipku j, ako ponudeni niz slova JEST postojeća riječ u njemačkom jeziku.</p> <p>Pritisnite tipku f, ako ponudeni niz slova NIJE postojeća riječ u njemačkom jeziku.</p> <p>Za nastavak pritisni tipku razmak.</p>
<p>Molim Vas da postavite desni kažiprst na tipku j, a lijevi kažiprst na tipku f.</p> <p>Čim pritisnete tipku j ili f, pojavit će se sljedeći niz slova.</p> <p>Molim Vas da budete što brži i što točniji.</p> <p>Kad budete spremni, pritisnite tipku j ili f i eksperiment može krenuti.</p>	<p>Kraj je prvoga dijela.</p> <p>U nastavku slijedi drugi dio eksperimenta.</p> <p>Struktura je jednaka prvom dijelu.</p> <p>Eksperiment kreće sa znakom fiksacije (točka).</p> <p>Za nastavak pritisni tipku razmak.</p>
<p>Nakon znaka fiksacije (točka) na ekranu bit će vrlo kratko predstavljena postojeća riječ u njemačkom jeziku, a zatim će biti predstavljen određen niz slova.</p> <p>Neki nizovi slova predstavljat će postojeće riječi (imenice) hrvatskoga jezika, a neki će predstavljati nepostojeće riječi hrvatskoga</p>	<p>tržnica</p>

<p>jezika.</p> <p>Slijede primjeri za nizove slova koje predstavljaju postojeće i nepostojeće riječi u hrvatskom jeziku.</p> <p>Za nastavak pritisni tipku razmak.</p>	
--	--

<p>Ili ako se radi o nepostojećoj riječi u hrvatskom jeziku:</p>	<p>melat</p>
--	--------------

<p>Vaš je zadatak ŠTO JE BRŽE MOGUĆE odlučiti je li ponuđeni niz slova riječ u HRVATSKOM jeziku ili nije.</p> <p>Mjeri se Vaša reakcija na drugu, hrvatsku riječ.</p> <p>Pritisnite tipku j, ako ponuđeni niz slova JEST postojeća riječ u hrvatskom jeziku. Pritisnite tipku f, ako ponuđeni niz slova NIJE postojeća riječ u hrvatskom jeziku.</p> <p>Za nastavak pritisni tipku razmak.</p>	<p>Molim Vas da postavite desni kažiprst na tipku j, a lijevi kažiprst na tipku f.</p> <p>Čim pritisnete tipku j ili f, pojavit će se sljedeći niz slova.</p> <p>Najprije ćete napraviti nekoliko zadataka za vježbu.</p> <p>Molim Vas da budete što brži i što točniji.</p> <p>Kad budete spremni, pritisnite tipku j ili f i vježba može krenuti.</p>
--	---

<p>Ovo je bila vježba.</p> <p>U nastavku ćemo prijeći na eksperiment.</p> <p>Trebat ćete učiniti isto kao i do sada.</p> <p>Za nastavak pritisni tipku razmak.</p>	<p>Vaš je zadatak ŠTO JE BRŽE MOGUĆE odlučiti je li ponuđeni niz slova riječ u HRVATSKOM jeziku ili nije.</p> <p>Mjeri se Vaša reakcija na drugu, hrvatsku riječ.</p> <p>Pritisnite tipku j, ako ponuđeni niz slova JEST postojeća riječ u hrvatskom jeziku.</p> <p>Pritisnite tipku f, ako ponuđeni niz slova NIJE postojeća riječ u hrvatskom jeziku.</p>
--	---

	Za nastavak pritisni tipku razmak.
--	------------------------------------

Molim Vas da postavite desni kažiprst na tipku j, a lijevi kažiprst na tipku f.	
---	--

Čim pritisnete tipku j ili f, pojavit će se sljedeći niz slova.	
---	--

Molim Vas da budete što brži i što točniji.	
---	--

Kad budete spremni, pritisnite tipku j ili f i eksperiment može krenuti.	
--	--

	Ovo je kraj eksperimenta.
--	---------------------------

	Zahvaljujem na sudjelovanju.
--	------------------------------

16.3. Popis parova ciljnih riječi i pripadajućih usmjerivača u oba eksperimenta

Popis ciljnih riječi i usmjerivača koji su povezani s ciljnom riječi i koji to nisu u prvom eksperimentu. Ciljna je riječ na njemačkom, dok je usmjerivač na hrvatskom.

Ciljna riječ	Povezan usmjerivač	Nepovezan usmjerivač
Fisch	pastrva	zaslon
Elefant	surla	naručje
Küche	jelo	paperje
Fenster	pogled	preljub
Ohr	sluh	govedo
Gras	livada	spremnik
Chor	crkva	orah
Zucker	sol	natpis
Wäsche	pranje	pokrov
Ring	vjenčanje	duda
Hund	prijatelj	letva
Gurke	vrhnje	tenisica
Schwester	brat	barka
Soldat	vojska	petlja
Tür	kvaka	krik
Schaf	vuna	porotnik
Fliege	kukac	kresta
Biene	osa	bunar
Körper	čovjek	novak
Ausgang	ulaz	melem
Schule	đak	udica
Muskel	snaga	truba
Kohl	salata	topot
Erdnuss	maslac	sušilo
Gesicht	osmijeh	brusnica
Eis	kocka	nastava
Blut	ozljeda	blagajna
Vogel	let	kesten
Wolke	kiša	kopito

Socke	cipela	siročje
Kopf	mozak	konj
Mütze	šešir	maćeħa
Haut	jakna	pladanj
Schneider	odijelo	umak
Spitzer	olovka	mornar
Stein	krš	odrezak
Ofen	drvo	pod
Dach	crijep	meduza
Brücke	rijeka	palača
Sand	plaža	kos

Popis ciljnih riječi i usmjerivača koji su povezani s ciljnom riječi i koji to nisu u prvom eksperimentu. Ciljna je riječ na hrvatskom, dok je usmjerivač na njemačkom.

Ciljna riječ	Povezan usmjerivač	Nepovezan usmjerivač
majka	Vater	Rad
ruža	Dorne	Koch
juha	Löffel	Kunde
kuća	Dach	Brief
vuk	Wald	Tasche
mlijeko	Kuh	Lupe
vatra	Brand	Seife
djed	Oma	Markt
polje	Bauer	Essig
žena	Mann	Tomate
stol	Stuhl	Delfin
kruh	Butter	Becher
ruka	Finger	Hütte
limun	Frucht	Stange
zlato	Schmuck	Deckel
stroj	Arbeiter	Adler
hlače	Jeans	Lachs
mrkva	Hase	Luft
cvijet	Duft	Helm

vjetar	Sturm	Reis
jaje	Huhn	Leber
sir	Löcher	Lehrer
lav	Mähne	Mauer
zub	Mund	Glas
pivo	Schaum	Felsen
nož	Gabel	Zaun
kanta	Müll	Thron
kruška	Obst	Zeitung
konac	Nadel	Kasse
bačva	Wein	Arzt
vrat	Kette	Stock
jastuk	Bett	Ball
kruna	König	Tüte
pramen	Haar	Boot
mlin	Mehl	Stimme
čavao	Hammer	Teller
polica	Buch	Kalb
sanjke	Schnee	Schere
slama	Heu	Niere
šalica	Tee	Müller

Popis pseudoriječi s pripadajućim usmjerivačima. Usmjerivač je na njemačkom odnosno hrvatskom. Pseudoriječ odgovara fonotaktičkim pravilima hrvatskoga odnosno njemačkoga jezika. Pseudoriječi su jednake u oba eksperimenta.

Usmjerivač (njemački)	Pseudoriječ	Usmjerivač (hrvatski)	Pseudoriječ
Ziege	vuba	sestra	Schlein
Daumen	zmaki	nož	Rekte
Stirn	spoglik	mreža	Lystel
Keller	pacofa	čamac	Seuler
Wagen	zamovitik	čvor	Nezefe
Kunde	ožniča	šator	Zwächen
Eltern	prndiki	ulje	Zöppen
Knödel	žnjun	ujak	Knaf

Tante	poc	škruga	Takin
Speck	lečavo	britva	Phlede
Braut	selarpra	mast	Rühne
Dampf	vitoneća	potpis	Putflug
Täter	povopšte	češalj	Deise
Gewürz	resun	svjedok	Drapfle
Bart	giza	polica	Quotte
Darm	ustuć	vozač	Kando
Drucker	flaja	kuhar	Kolf
Richter	siniljur	stric	Nigesker
Laden	svrajarce	očuh	Enafös
Glocke	brcivo	dvorac	Blafe
Klingel	krod	ured	Dosse
Spargel	ukrija	kazalište	Pfleches
Messe	rilna	živac	Kirle
Tempel	zaljoć	kovčeg	Schaspuse
Tuch	omušiš	kip	Napelle
Strumpf	čkan	stup	Pleield
Nebel	šerda	gusar	Hittes
Gürtel	olerma	oluja	Strirt
Kneipe	prošnevo	list	Geflalst
Abfall	bišor	dvorište	Herzam
Lager	pemimniva	luka	Sährer
Welle	puroč	povez	Katglut
Dose	ladupa	kist	Istest
Tulpe	borbun	pozornica	Jorlage
Pass	pozjage	gost	Sänntam
Kasse	onavga	hrđa	Rorf
Schnaps	otprika	ocat	Salp
Wurzel	uzmog	tajnik	Tamter
Ameise	svežurg	put	Tätte
Sack	zaža	vino	Brickt

Popis ciljnih riječi i usmjerivača koji su povezani s ciljnom riječi i koji to nisu u drugom eksperimentu. Ciljna je riječ na njemačkom, dok je usmjerivač na hrvatskom.

Ciljna riječ	Povezan usmjerivač	Nepovezan usmjerivač
Fisch	kornjača	zaslon
Elefant	majmun	naručje
Küche	soba	paperje
Fenster	ključanica	preljub
Ohr	grlo	govedo
Gras	voda	spremnik
Chor	ples	orah
Zucker	škrob	natpis
Wäsche	sude	pokrov
Ring	dar	duda
Hund	dječak	letva
Gurke	grah	tenisica
Schwester	nećakinja	barka
Soldat	policajac	petlja
Tür	kapija	krik
Schaf	koza	porotnik
Fliege	mačka	kresta
Biene	matica	bunar
Körper	truplo	novak
Ausgang	ulaz	melem
Schule	bolnica	udica
Muskel	prijelom	truba
Kohl	krumpir	topot
Erdnuss	lješnjak	sušilo
Gesicht	ogledalo	brusnica
Eis	sloj	nastava
Blut	tekućina	blagajna
Vogel	zmija	kesten
Wolke	sunce	kopito
Socke	cipela	siročje
Kopf	noga	konj
Mütze	majica	maćeha
Haut	odjeća	pladanj

Schneider	prodavač	umak
Spitzer	olovka	mornar
Stein	mač	odrezak
Ofen	šibica	pod
Dach	zid	meduza
Brücke	brod	palača
Sand	tlo	kos

Popis ciljnih riječi i usmjerivača koji su povezani s ciljnom riječi i koji to nisu u drugom eksperimentu. Ciljna je riječ na hrvatskom, dok je usmjerivač na njemačkom.

Ciljna riječ	Povezan usmjerivač	Nepovezan usmjerivač
majka	Vater	Rad
ruža	Lilie	Koch
juha	Nudeln	Kunde
kuća	Wohnung	Brief
vuk	Fuchs	Tasche
mlijeko	Honig	Lupe
vatra	Flamme	Seife
djed	Oma	Markt
polje	Wiese	Essig
žena	Mann	Tomate
stol	Stuhl	Delfin
kruh	Fleisch	Becher
ruka	Finger	Hütte
limun	Saft	Stange
zlato	Silber	Deckel
stroj	Gerät	Adler
hlače	Hemd	Lachs
mrkva	Kartoffel	Luft
cvijet	Garten	Helm
vjetar	Brise	Reis
jaje	Küken	Leber
sir	Schinken	Lehrer
lav	Bär	Mauer

zub	Gebiss	Glas
pivo	Schnaps	Felsen
nož	Dolch	Zaun
kanta	Flasche	Thron
kruška	Apfel	Zeitung
konac	Schnur	Kasse
bačva	Kiste	Arzt
vrat	Brust	Stock
jastuk	Decke	Ball
kruna	König	Tüte
pramen	Haar	Boot
mлин	Hof	Stimme
čavao	Hammer	Teller
polica	Schrank	Kalb
sanjke	Schnee	Schere
slama	Heu	Niere
šalica	Tee	Müller

16.4. Popis tablica

Tablica 1: Popis od nekoliko naslova članaka koji su korišteni kao primjer prikaza funkcioniranja LSA-a.....	26
Tablica 2: Matrica tipa ($m \times n$) u kojoj redci (m) predstavljaju riječi koje se mogu nalaziti u naslovima znanstvenih članaka, a stupci (n) naslove znanstvenih članaka.....	28
Tablica 3: Rastav na singularne vrijednosti matrice tipa (15×10) iz Tablice 2.....	33
Tablica 4: Matrica A_k nastala kao umnožak triju matrica U_k , Σ_k i V_k iz Tablice 3.....	37
Tablica 5: Interkorelacije (r) kontekstnih vektora između računalno neobrađenih podataka (gore) i između podataka dobivenih rekonstrukcijom matrice smanjene dimenzionalnosti (dolje).....	38
Tablica 6: Matrica prema modelu HAL.....	41
Tablica 7: Primjeri parova usmjerivača i ciljne riječi u eksperimentu s diskretnim asocijacijama.....	87
Tablica 8: Prosječan broj slova po riječima u eksperimentu s diskretnim asocijacijama.....	87
Tablica 9: Mjere deskriptivne statistike u prvom eksperimentu za vrijeme reakcije s obzirom na povezanost ciljne riječi i usmjerivača te jezik na kojem je zadana ciljna riječ ($N = 54$).....	90
Tablica 10: Primjeri parova usmjerivača i ciljne riječi u eksperimentu s ispitnim materijalom generiranoga pomoću računalnoga programa <i>word2vec</i>	100
Tablica 11: Prosječan broj slova po riječima u eksperimentu s ispitnim materijalom generiranoga pomoću računalnoga programa <i>word2vec</i>	100
Tablica 12: Mjere deskriptivne statistike u drugom eksperimentu za vrijeme reakcije s obzirom na povezanost ciljne riječi i usmjerivača te jezik na kojem je zadana ciljna riječ ($N = 44$).....	103

16.5. Popis slika

Slika 1: Rastav na singularne vrijednosti matrice A tipa ($m \times n$) ranga r na tri matrice ranga k	31
Slika 2: Dvodimenzionalni prikaz rasporeda riječi i kontekstā dobiven iz rastava na singularne vrijednosti na temelju podataka iz Tablice 3.....	35
Slika 3: Prikaz kuta između riječi i kontekstā.....	36
Slika 4: Prikaz konvolucije za sve n-grame koji okružuju riječ <i>kaznenu</i> iz rečenice "Zadali su mi kaznenu zadaću" (Lenz, 1999: 9).....	50
Slika 5: Model CBOW i model skip-gram.....	53
Slika 6: Inhibicijski kontrolni model (IC).....	66
Slika 7: Bilingvalni interakcijski model leksičkoga pristupa (BIMOLA).....	68
Slika 8: Revidirani hijerarhijski model (RHM).....	71
Slika 9: Hipoteza o epizodičkom pamćenju.....	73
Slika 10: Bilingvalni interaktivni aktivacijski model (BIA).....	75
Slika 11: Model semantičke, ortografske i fonološke interaktivne aktivacije (SOPHIA).....	77
Slika 12: Bilingvalni interaktivni aktivacijski model plus (BIA+).....	78
Slika 13: Razvojni model bilingvalne interaktivne aktivacije (BIA-d).....	79
Slika 14: Model distribucijskih konceptualnih obilježja (DCFM).....	80
Slika 15: Model značenja reprezentacija dvojezične memorije.....	82
Slika 16: Prikaz koncepta.....	86
Slika 17: Tijek eksperimenata.....	89
Slika 18: Prosječne vrijednosti vremena reakcije u prvom eksperimentu s obzirom na povezanost ciljne riječi i usmjerivača.....	91
Slika 19: Prosječne vrijednosti vremena reakcije u prvom eksperimentu s obzirom na jezik na kojem je zadana ciljna riječ.....	91
Slika 20: Prosječne vrijednosti vremena reakcije u drugom eksperimentu s obzirom na povezanost ciljne riječi i usmjerivača.....	103
Slika 21: Prosječne vrijednosti vremena reakcije u drugom eksperimentu s obzirom na jezik na kojem je zadana ciljna riječ.....	104

17. ŽIVOTOPIS AUTORA

Lukas Paun rođen je 1986. godine u Fuldi (SR Njemačka) gdje je pohađao niže razrede osnovne škole. 1996. godine seli se u Požegu gdje nastavlja školovanje i završava tamošnju gimnaziju. Sve je razrede gimnazije prošao s odličnim uspjehom zbog čega je oslobođen polaganja maturalnoga ispita. 2000. godine upisuje Filozofski fakultet Sveučilišta u Zagrebu, smjer germanistika i kroatistika. Tijekom studija pokazao je zanimanje za dijalektologiju slavonskoga kraja. Pod vodstvom prof. dr. sc. Mire Menac-Mihalić piše seminarski rad *Fonološka i morfološka analiza govora Ivandola* za koji 2007. godine dobiva Rektorovu nagradu. Diplomski rad pod nazivom *Naturwissenschaften als Vorbild für die Sprachwissenschaft im 19. Jahrhundert (Prirodne znanosti kao uzor jezičnoj znanosti u 19. stoljeću)* piše pod mentorstvom prof. dr. sc. Maje Anđel za vrijeme studijskoga boravka na Sveučilištu u Erfurtu kamo odlazi kao stipendist DAAD-a (njem. *Deutscher Akademischer Austauschdienst* - engl. *German Academic Exchange Service*) 2009. godine. Odmah po povratku u domovinu 2010. brani diplomski rad.

Školske godine 2010./2011. zapošljava se kao učitelj njemačkoga jezika u OŠ fra Kaje Adžića u Pleternici gdje predaje nižim i višim razredima sve do danas. Od školske godine 2013./2014. prelazi djelomično na OŠ Julije Kempfa u Požegu gdje također radi do danas. 2011. godine položio je stručni ispit za zanimanje učitelja njemačkoga jezika. Redovito se stručno usavršava na skupovima koje organizira Agencija za odgoj i obrazovanje te Goetheov institut. U sklopu angažmana u Goetheovu institutu boravi na stručnim skupovima u Srbiji i Njemačkoj. Godine 2018. dobiva stipendiju za učitelje japanske vlade. U Japanu boravi godinu i pol dana usavršavajući se u svom pedagoškom radu na Sveučilištu u Nagoyi (prefektura Aichi) te na Sveučilištu u Oostu (prefektura Shiga). Vraća se u Hrvatsku 2020. godine i nastavlja s pisanjem disertacije i radom u osnovnom obrazovanju.

Radovi:

Paun, Lukas (2022). Utjecaj semantičke povezanosti na razumijevanje tekstova za čitanje i slušanje. *Strani jezik* 51(2): 257-282.