Filozofická fakulta Univerzity Palackého v Olomouci

Katedra obecné lingvistiky

Evidence for regular occurrence of low-frequency items and its implications for mental lexicon models

magisterská diplomová práce

Autor: Bc. Zdeněk Joukl

Vedoucí práce: doc. Mgr. Dan Faltýnek, PhD.

Olomouc

2022

Prohlášení

Prohlašuji, že jsem magisterskou diplomovou práci „Evidence for regular occurrence of low-frequency items and its implications for mental lexicon models“ vypracoval samostatně a uvedl jsem veškerou použitou literaturu a veškeré použité zdroje.

V Olomouci dne Podpis

Abstrakt

Název práce: Evidence pro pravidelný výskyt nízkofrekvenčních položek a její důsledky pro modely mentálního lexikonu

Autor práce: Bc. Zdeněk Joukl

Vedoucí práce: doc. Mgr. Dan Faltýnek, PhD.

Počet stran a znaků: 49 stran, 140 000 znaků

Počet příloh: 10

Abstrakt (minimálně 900 znaků): Práce rozebírá koncept superhapaxů, tedy nízkofrekvenčních lexikálních položek, které mají u jednotlivých autorů pravidelný výskyt přibližně jednou každých 6250 slov a zdají se být vázané na téma, autoři mají svá preferovaná témata, o kterých často mluví. Zdá se, že priming tématu zvyšuje aktivaci těchto položek. Práce evidenci vyhodnocuje ve vztahu k představám o mentálním lexikonu. Zásadní je kladení důrazu na individuální strukturaci mentálního lexikonu na základě naučeného kontextu spolu s některými slovy. Jedním z možných vysvětlení tvorby hapaxů v mentálním lexikonu může být koncept S-P a S-P-S posuvu, tedy preference v syntagmatických nebo paradigmatických asociacích, kdy se s dospíváním určité výrazy sémanticky omezují a jejich kontextové užití se solidifikuje. Práce se snaží nalézt modely, které by co nejlépe byly schopny akomodovat tuto novou evidenci a současně zkritizovat modely, které takové evidenci nevyhovují. Jako zvláště zajímavé se jeví model multiplexový a modely vícejazyčné. Nakonec práce přináší data z analýzy sentimentu superhapaxů v italském jazyce.

Klíčová slova: hapax, superhapax, mentální lexikon, model, frekvence, kontext

Abstract

Title: Evidence for regular occurrence of low-frequency items and its implications for mental lexicon models

Author: Bc. Zdeněk Joukl

Supervisor: doc. Mgr. Dan Faltýnek, PhD.

Number of pages and characters: 49 pages, 140 000 characters

Number of appendices: 10

Abstract (900 characters): The paper discusses the concept of superhapaxes, i.e. low-frequency lexical items that have a regular occurrence of approximately once every 6250 words for each author and seem to be tied to a topic; authors have their preferred topics that they often talk about. Topic priming seems to increase the activation of these items. This paper evaluates the evidence in relation to ideas about the mental lexicon. The emphasis on individual structuring of the mental lexicon based on learned context along with some words is crucial. One possible explanation for the formation of hapaxes in the mental lexicon may be the concept of S-P and S-P-S shift, i.e., a preference in syntagmatic or paradigmatic associations, whereby certain terms become semantically restricted and their contextual use becomes solidified as they mature. This paper seeks to find models that are best able to accommodate this new

evidence, while critiquing models that do not accommodate such evidence. The multiplex model and multilingual models appear to be particularly interesting. Finally, the paper presents data from a sentiment analysis of superhapaxes in Italian.

Keywords: hapax, superhapax, mental lexicon, model, frequency, context

Summary

Introduction

The aim of this thesis was to evaluate the evidence from experiments with hapaxes, on which the concept of superhapax (simply put, rare word that we use frequently) is based, in its relation to ideas about the mental lexicon, and to try to find those models that are built in such a way that they can explain the existence and properties of superhapaxes and, on the contrary, to criticize those models that cannot accommodate this new evidence.

It begins with a categorical overview of the typology of hapaxes and their roles in language and the mental lexicon. Based on experiments, Faltýnek and Matlach (2021) found that people use certain low-frequency items repeatedly (these items are called superhapaxes). This may be related to their preferences to talk about certain topics. Such a finding has important implications for ideas about the mental lexicon and how speech is constructed, because hapaxes have a cohesive role (they include both content and function words).

It also gives an overview of the most relevant findings from the literature, which must be taken into account when studying the mental lexicon and its functioning. Many of the models focus only on certain aspects of language processing and are not based on psychological findings about memory, although they are supported by interesting experimental evidence. Some hypotheses may be based on bad assumptions, which is why they are partly confirmed and partly not confirmed. Thus, the thesis will provide a list of implications for particular ideas about the mental lexicon in light of the superhapax evidence, and a preliminary attempt at a model of its own will be included.

In the practical part, new data from the Italian language will be supplied and analyses will be carried out to support theoretically what we assume on the basis of experiments already carried out on different languages. A small-scale sentiment analysis will be conducted to look for significant results of sentiment and topic attraction of superhapaxes.

Hapaxes: low-frequency lexical items

Before we can start working with hapaxes, we need a good (and one that is as precise as possible) definition of this concept. The Greek expression *hapax legomenon* means "what is said only once," which corresponds with the traditional conception of hapax legomenona in linguistics, namely that of a word occurring only once in a given text. However, we are entering a new area of research, within which we will constitute a whole new set of different conceptions of hapaxes. Therefore, the term "hapax" can be at times quite misleading. The object of our interest could be described also more generally as "low-frequency words."

This work is concerned with hapaxes or low-frequency lexical items (and in particular some of their types) in relation to different models of mental lexicon. The conception of hapaxes must be adequately adjusted to this context. To answer the question of what actually is a hapax in the mental lexicon, we must at first explain what mental lexicon is and also describe all the phenomena that can be called hapaxes.

In diachronic disciplines, a hapax is often an error: a misspelled word in a text. The identification of a hapax can be sometimes tricky, especially when examining ancient manuscripts written in a language that is not so well-known or understood. In such cases, the researcher can erroneously label an infrequent word as an error, despite the word only having a low (or no) occurrence in the preserved texts of the language. We cannot confidently say that

this particular word was not used in spoken communication or that it did not occur in any other text.

In the context of cognitive sciences and neurosciences, the term "hapax" requires yet further specification. When we work with textual objects, more exact definitions are possible than during the work with cognitive objects that are difficult to grasp. A hapax in a mental lexicon can be a production error. If we are to store a certain word in our mental memory, we must be able (although we are not always able) to access it, which means its occurrence cannot be as low as 1, but rather higher.

When reading, we could come across a certain word only once and never think of it anymore. This could be called an *absolute hapax* (also) in the context of mental lexicon.[1]

Later, we will see that there are many details which may influence our way of defining hapaxes. For instance, there is a lot of evidence for morphological processing of words in the brain, which means we might not be storing every individual word form. There are many other ways in which our brains might be processing the language, which is yet another reason why it is problematic to define hapax in the mental lexicon. It might be necessary to define hapax in the context of every individual model or a group of models.

Considering all the previously mentioned matters, we end up needing rather a generalized definition of hapaxes to start our work: "a low-frequency item." In certain senses and contexts, the traditional definition will be preserved: for instance, in their analyses, Faltýnek and his team are working with pre-cut texts that contain items with only one occurrence in a given textual segment, but not within the whole text (and so the items with frequency 1 or slightly higher within a particular cut are considered hapaxes, although they may reappear in the other segments of the text).[2]

At first, before approaching and considering every individual model, we should create a list – or a taxonomy of what can be considered a hapax.

Categorization of hapaxes

Produced vs perceived

The first basic possible division of hapaxes we can make is separating items that emerge during speech/text production and those that emerge during speech/text reception.

Production hapaxes are produced by the speaker/writer in a certain way that can further our taxonomy: it can be intentional or unintentional misspelling of a handwritten or computer-written word, the individual can invent a new expression purely in his mind (mere thought) or during communicational interaction with other speakers. This word or expression can be reused by the inventor a certain number of times, maybe also by other people who had become acquainted with it.

Reception hapaxes are hapaxes heard or read by the listener or the reader, who then processes them further and pays more or less attention to them. It can be a word in a book the reader read (or other kind of text) or a word used by someone the individual listened to.

[1] Already defined by Faltýnek (2021)

[2] As we will see later, this type of hapaxes is called "superhapax."

Absolute vs local

Quantitatively speaking, we can distinguish between absolute and local hapaxes. Still in this context, it is necessary keep in mind that a certain expression can be a hapax within a concrete book, but not necessarily within the whole body of usage of its author. An absolute hapax can be an item with frequency as high as 1, while the extent of the textual object of appearance of this hapax remains undefined. Thus, an item can be considered an absolute hapax for a particular section of a text, an absolute hapax for a particular text or an absolute hapax for the whole of the texts of one author (i.e. this author would have used it only once in his/her life).

Phonological

We may also want to step out of the traditional level of language considered for the research of hapaxes (i.e. the lexical one) and admit the possibility of considering them on all levels of language.

The phonological hapax will be a complicated one to delimit. It could mean a specific place or manner of articulation that occurs with low frequency (as an authorial feature or a sequence-specific phenomenon or an articulation slip), or it could be a variational feature that can be used to identify the speaker. Such item could occur regularly due to anatomical or habitual predispositions of the individual speaker.

A precise delimitation of phonological hapaxes is, however, not necessary for the purposes of this thesis. Therefore, we will leave this area for later theorization. The analyses of Faltýnek & Matlach (2021) focus mainly on lexical hapaxes, because this traditional category remains the easiest and the most effective one to analyze and to be used for authorship attribution.

Morphological

A very interesting category are hapaxes on the morphological level. Further research will be needed to clarify their role in language, textual cohesion, authorial speech/text production etc. Different mental lexicon models work with morphology to different extent and there is not enough certainty in what our brain or mind really does on the morphological level during language processing. However, we will need to elaborate on this category, because it could have significant implications for the processing of lexical hapaxes.

Concretely speaking, morphological hapaxes can be simply viewed as low-frequency morphemes. However, they must be considered in context, at least in some cases. If a speaker/writer prefers a particular form or variant of a word or a group of words, a morphological hapax can be identified (Czech "balíčcích/balíčkách"). Such hapaxes could sometimes also emerge due to – normatively speaking – erroneous analogy (Czech "s hranolkami" instead of s "hranolky").

Another possibility is looking at the langue and identifying rare phenomena occurring due to onomatopoeia (since they often contain untypical sounds) or phonestemes (although these are submorphemic). An individual may have a preferred (or may be used to) morphological way of expressing certain emotions.

Lexical

The traditional category is dependent on our definition of the word or lexical item. Lexical hapaxes will be the focus of this thesis, since they seem to be the "default" unit of organization at the base of most neural language processing.

Syntactic

On the syntactic level, one might speak of low-frequency word combinations equal to bi-grams or larger n-grams, but also skip-grams. A certain unusual word combination may belong to the typical repertoire of a certain author. Since there are important implications of syntactic processing for the lexical level, we must take into consideration also the syntactic hapaxes and look at some models of syntactic processing in the brain/mind.

When reading the Czech translation of John Steinbeck's Of Mice and Men, we frequently encounter otherwise unusual expressions such as "bez obalu" ("he said it outright") or "pospíchal ze dveří" ("he was rushing out the door"). Steinbeck frequently uses also some unusual onomatopoetic expressions such as "řachtění talířů" ("clattering of plates") or an unusual comparison "jako přišlápnutý červ" ("like a squashed worm").

Semantic

For the semantic level, we must specify the level of uniqueness. The delimitation of semantic hapaxes is dependent on our view of meaning and context. It can be said that there are only contextual meanings. In such case, however, everything would be a hapax. This is not useful for our work at all (we will work with semantic hapaxes to a minimal extent only though). A slightly reduced extent of application would mean that hapaxes are the necessary precondition of creativity standing at the base of scientific advancement and new ideas. A semantic hapax would be also every metaphorical usage of a word in e.g. poetry. A more suitable approach would be using the hapax label for items used in their low-frequency meaning (that may or might not occur in a dictionary listing of the meanings of this word).

Examples of semantic hapaxes can be the terms "čep" and "oddenek" used by an art historian in the context of architectonic columns as an analogy to botanic terminology. "Čep" is the area (cross-section) in the upper part of a trunk of a tree, while "oddenek" is the lower part (rhizome). Tree trunks resemble architectonic columns, which inspires this rare analogy.

Error

From another point of view, we should single out also hapaxes that are actually production errors. Such errors can sometimes become lexicalized, especially within idiolects ("obranismus," "diskuzovat," "pododělat" and "maturika" are some of the examples of words that were at some point used by our colleagues due to their slip of the tongue, but they found them appealing and reused them later, they became an inbuilt part of their mental lexicons with low frequency of usage). They are explained later in the Economizing function chapter.

Error hapaxes can receive different amounts of attention: if it is an uninteresting and common slip of the tongue, we usually pay no attention to it. Sometimes it remains in our short-term memory and if we pay a lot of attention to it, it may become a part of our mental lexicon and of the long-term memory.

Somewhat relevant to understanding such error hapaxes is the phenomenon of parapraxis, first described by Freud in the first chapter of his Psychopathology of Everyday Life (1901), also called the Freudian slip: Freud wanted to say the name of the painter of the frescoes in Orvieto Cathedral, but could not remember exactly. So he said the painter's name was something like Botticelli or Boltraffio, when in fact it was Signorelli. The interpretation of why he remembered these two names is highly contested, but it does suggest the ways in which information can flow and how it is stored and accessed – and how errors can occur. Freud says he probably associated it with the village of Trafoi, where he had earlier received a report of the suicide of one of his

patients. He had associated the village with a topic he had discussed shortly before. He reinterpreted signore in Signorelli to German Herr, which in turn he associated with BOsna and HERcegovina, where he had gone when he had a conversation on a topic related to Trafoi. Processes like this lead us to the idea of approximation based on contextually available information, of the brain trying to make the best approximation of some idea that is no longer attainable but we have some leftover information around. This suggests to us the attractiveness of the mental lexicon model based on prototypes, which will be described later.

Parapraxis is also the basis for a series of research studies investigating not only errors but also repair processes. Sometimes thinking is faster than speaking and we mispronounce something, but at the same time we have self-monitoring mechanisms that allow us to feedback to ourselves and respond to the error by, for example, correcting or apologizing to the audience. These self-monitoring mechanisms have been investigated, for example, by Levelt (1983, 1989), Blackmer & Mitton (1991), van Wijk & Kempen (1987) or Lashley (1951). Self-monitoring mechanisms would suggest some intuition that the existence of approximation mechanisms into prototypes might give us. We have experienced certain prototypical norms, and when we inadvertently violate them, we often notice, or say something in a foreign language that we are not sure about and notice that it does not sound quite right. But at the same time, prototypes can make us blind to errors, namely in reading: we often do not notice a misspelled word at first glance because we expect a certain word in a given context, and when it resembles the expected word, we automatically evaluate it as that word. It is only on further reading that we notice that it is misspelled.

Now the question, still related to error-making is whether superhapaxes (see the definition below) are somehow related to errors and hesitation and why. It could be that low-frequency words are more error-prone due to their lack of solidification, fewer synapses, etc., yet authorial superhapaxes should have an advantage. They might constitute important network nodes.

Hesitation is discussed by Lerner (2013), who creates their categorization. According to him, hapaxes can be divided into: quieter elements that speakers are afraid to say; laughter, obscuring some words; euphemism; difficulty of finding a word in the mental dictionary (in this situation, other conversational participants can help with finding it); the interjection uhm/ahm or prolongation of the previous word; compositional projects; attempts to complete (e. g. a sentence) by producing a precise formulation; hesitation before a delicate word; thinking about the possibility of a different completion; suspension (e.g. to insert an emotive sound, etc.), social solidarity; or thinking about the sequential arrangement of delicate formulations.

During the examination of Berlusconi's hapaxes, we came across hapaxes that might fall into the last category, like the superhapax "vero," which often seems to have a somewhat padding role for Berlusconi.

Schlegloff (2013) further describes the corrective operations that are performed when an error occurs. But they are rather syntactic, and this implies that there must be macro-planning, our speech follows a certain trajectory and its invention precedes the utterance of the speech. More relevant to the operating of the mental lexicon is the phenomenon of parapraxis rather than these corrective operations.

Idiolectal/sociolectal differentiation

Some low-frequency words can be highly frequent within specific contexts. This regards terminology, but also sociolects. Some words can have unusual variants that can be preferred by certain social groups or individuals, their usage may be context-dependent ("má"/"moje" – i.e. "my;" "tofužel"/"bohužel" – as used in the Czech vegan sociolect with the meaning of "unfortunately").

Multilingual speaker's hapaxes

Language learners can be a very interesting source of hapaxes. A person learning a foreign language will begin to use words from their target language also when speaking the other languages they know, e.g. because they will consider them more apt or accurate for a certain context. Appropriate use of low-frequency items in a suitable context may indicate higher lexical sophistication and a higher level of foreign language (and indeed also L1) acquisition.

However, the research of multilingual superhapaxes is quite complicated because it involves interweaving layers of the mental lexicon and there is a need to separate the samples well, and these samples are then taken out of context if the speaker switches languages. Moreover, the use of a particular language is likely to prime the speaker to that language and create a bias to other languages. At the very least, even such research will be useful for determining the transferability of superhapaxes across languages.

Superhapaxes

Faltýnek (based on Faltýnek & Matlach 2021) introduces a new concept (and a crucially important one) of superhapax, which is an item appearing in different authorial texts. This item has low frequency, but appears regularly in different texts/performances of the author. It means they might be found in the center of sparse networks constituted of around one thousand words. Although in the course of time individuals will have very immense vocabulary changes (down to reading influential books or entering different social groups or ageing), superhapaxes are items that remain. They represent the author in terms of psychological profiling: it epitomizes germane traits of their character. It may be related to some details the person pays attention to (I have a friend whose favorite color is pistachio and he is able to find it anywhere – and after some time he always points at something pistachio; others focus rather on shapes or sizes) or to the way the person uses some abstract verbs. Faltýnek and his team identified the superhapax "winter" in Kerouac's work: it might be because the author kept feeling cold and on that account he traveled to warmer countries.

Superhapaxes are the most relevant class of hapaxes for our commenting of the models of mental lexicon, because they seem to be a language-constituting cohesive element that holds together the lexical networks. The analysis of superhapaxes will have an impact on the models of mental lexicon and on the studies of priming.

Based on superhapaxes identified in text segments of 6250 words (as in Faltýnek & Matlach 2021), 100% author identification is assured, with Matlach's 100% clustering based on all but the last case of French. To achieve clustering everywhere, we need the limit of 3500 hapaxes, with which the values for F1, precision and recall take 100%.

More general, semiotic hapaxes

There is also the possibility of taking into account the existence of hapaxes outside of natural language and developing a more general semiotic theory. After all, let us recall one meme circulating on the Internet (2018):

What do you call a lego piece that appears only once in a given set? – A hapax legomenon.

– Harrison Lemke

In virtually any semiotic system, one can consider the concept of a hapax, a unit that will appear only once in some specified section or set of elements. It will be different with the concept of a superhapax, which in natural language operates probably due to human cognitive properties tied to natural language, but is unlikely to work (at least not in the same way) in expressions other than just natural language (of course, this issue would need to be explored). One area that offers itself is the arts: e.g., artists regularly use an element that otherwise has a low frequency in their overall expression. As an example, consider the willow in the works of the Italian Baroque painter Domenico Fetti.

Role of hapaxes

Information theory

Hapaxes and superhapaxes will have a role in information theory, and this issue is addressed by Finn (1977-78), in whose work we find interesting insights. Low-frequency words carry more information, and low-frequency words that occur regularly in the text are related to the topic.

Finn considers that the processing of words with different frequencies depends on how many markers they have (i.e., meanings, grammatical forms, etc.). Rare words that are repeated in the text carry little information because they have exceptional transfer features. Finn mentions that Venezky & Calfee (1970) and LaBerge & Samuels (1974) work in their models with the observation that more frequent words are processed by readers with less effort than rarer words. The former write that when accessing high-frequency words, we enter a "highly organized" space that can therefore be searched quickly "on the basis of such features as initial letters and length. Less common words are stored differently and are more accessible by sound than any other form."

LaBerge & Samuels propose a model that assumes that words (depending on how much we have learned and mastered them) are processed with a certain degree of automaticity (high-frequency words will be processed automatically, we do not have to think about them deeply) and with a certain degree of accuracy, for which attention is crucial; low-frequency words require more attention to be processed.

Synergetic and quantitative linguistics

Since superhapaxes appear as an organizing element of the text, i.e. the topic determines the vocabulary used by the author, it would be worth reflecting on the description of the role of superhapaxes in Synergetic Linguistics. However, this is not possible due to the scope of this work.

When it comes to quantitative aspects and implications of hapaxes, the first question that comes to mind is the relation of hapaxes to Heap's Law (a statistic law of natural language texts to infinitely increase the size of their vocabulary with increasing length of texts). Fenxiang (2010) researched the hapax-vocabulary ratio on English texts and found a U-shaped dependency curve: at the beginning, until the text size reaches around 3 million words, the hapax-vocabulary ratio decreases and then it begins to increase and approach 1. The point of breakage of the curve will be different for different language types. Analytic languages are expected to have a lower

breakpoint than synthetic ones. Popescu and Altmann (2008) introduced the index of analytism, which can be used for researching the differences among different language types.

Economizing function

As concrete examples of certain efforts at economization that may have gone unnoticed or caused misunderstandings, I will cite the reports of several secondary school graduates: they experienced a slip of the tongue that led to pronouncing expressions, in some cases so hilarious that they lexicalized them. One of them used the term "obranismus" when she wanted to pronounce "obranný mechanismus" ("defense mechanism") when she got sick and wanted to complain about her poor immune system. In another context, she used the verb "diskuzovat" in the sentence "Klidně o tom můžeme diskuzovat" ("We are free to discuss this"), likening the word to the noun "diskuze" ("discussion"), so there was a hasty derivation of the verb in the context of the word "diskuze." The other graduate at the time, in a hasty speech, shortened the phrase "maturita z matematiky" ("secondary school graduation exam from maths") to one word: "maturika." He also has a lexicalized special verb form "pododělat," which uses a very unusual combination of prefixes and could be translated as "to do the finishing works." The third graduate eagerly talked about one of the books required reading for graduation, Karel Čapek's White Disease, which features a dictator called the Marshal. The graduate wanted to liken the Marshal to Hitler, but, primed by the Marshal at that moment, he called Hitler Mishler.

Poetic function

Hapaxes play an important role in poetry, where it is often necessary to slightly modify words to fit a rhyme or use unusual words.

Rhetoric function

Sometimes hapaxes are expressional means so specific that they attract attention. Many speakers or advertisements take advantage of this phenomenon to attract attention of customers. Even though someone would not otherwise buy the product, rhetorical devices including hapaxes can sound so appealing that they work. As an example, consider the excellent speaker known from Czech YouTube ads, Jiří Vokiel Čmolík, who uses expressions such as "pura vida" or "kulervoucí" ("amazeballs"), that are very unusual and sound very intense. The second one has much more power (with vulgar connotations) than its usual synonym "bombastický" ("grandiose").

Secret code

Hapaxes can be also thought of as a part of a secret code or a language of a small group of people (lovers teasing each other, summer camp participants, or a social group with specific interests).

Hapax grammar

When thinking about the grammatical roles of hapaxes, they appear across large distances and so the most relevant function to think of is the cohesive one. Nothing is known so far about how the hapaxes relate to the speech processing in the brain, to emotional and sentimental processing, whether they are related to the state of doubt, or to what extent they constitute text-forming elements. It seems that they are a cohesive element of different texts of one author. We note the compulsive repetition through which we can describe higher textual syntax. The relation of hapaxes to authorial production units (uninterrupted streams of consciousness) needs to be tested, they might also constitute transitions between subtexts of a particular text.

The most important implication might be that it is the hapax structure as a formal device, what organizes the text, i.e. top-down building from the semantic basis, sentences are interconnected through authorial hapaxes (Faltýnek 2021). They hint at what "properties of context the author prefers to express" and "which facts the author prefers to express." Both "lexicon and grammar are affected by this form of text structure," because superhapaxes include not only content words, but also functional words.

As Faltýnek argues, with 1000 autosemantics we have an overall thematic characteristic of the author. Every tenth word is one that the author likes to repeat, it is authorially obligatory, it controls the construction of sentences. We need to examine whether the author will show variations in word sentiment across history. If the variations are small, then this will be a very significant finding.

Sentiment

There is no doubt some words will have a specific emotional role. Consider, for example, a word that a person in love adopts from a loved one, e.g. because the loved one uses it often, even though it is otherwise unusual. Words could also have an unconscious emotional load, e.g. if they are learned (possibly as part of a specific environment) along with an emotionally loaded fact. For such cases, we are likely to associate a given word with a given emotional load or a given important fact, which could also be accessed more quickly through the use of the word in a different context (regarding the reaction time paradigm that will be further described later).

Given that hapaxes are infrequently used items, they might require more creative thinking and processing in the right hemisphere. Certain groups of words are known to be more related to the right hemisphere because of their affectedness. Typical examples of this can be vulgarisms, it might be also the language of lovers (for whom the name of the beloved other has a great affective significance, the same can be valid also for other unusual words typically used by one's beloved person) or words related to special interests of e.g. scientists, who will develop an emotional connection to the topic they are researching (every niche of science has its very special terminology, which becomes very frequently used by the researcher). However, the question still remains, whether positively connotated words are processed in the same way as negatively connotated words.

From the evolutionary point of view, Vakoch & Wurm (1997) give a very interesting hypothesis: "The organism can afford to take time processing information about emotions that reflect the conjunction of strength (high Potency), badness (low Evaluation), and slowness (low Activity), and this matched experimental results. However, when the words connote strength, badness, and quickness (high Activity), then the organism might be in peril, and much faster processing is called for. Again, this was consistent with our findings." They assume that emotional processing is sufficient through these few dimensions. In contrast, Hansen, McMahon & de Zubicaray (2019) argue that emotionally loaded words may take longer to respond to because the extra features steal attentional resources and act as distractors.

A question for our research, which would need to be tested with neuromethods, is whether superhapax words that have an emotional load have similar physiological responses to those of swearwords and taboo words in which the right hemisphere (thalamus) or amygdala is activated. If these are words associated with emotion (including the negative ones) or otherwise loaded, there may be a longer response to them because they act as distractors – features stealing attention resources (judged so but based on the picture-word interference paradigm, within

which the subjects are asked to name images of neutral objects meanwhile ignoring the accompanying distracting words).

As the hapax evidence shows, words have a closed sentiment neighborhood and are associated with a specific context in roughly approximately 70% of cases, there are intersections between superhapaxes and association with sensory content. For instance, he argues, the word "eating" in one of the authors analyzed would seem neutral, but in fact it was very contextual and triggered an experience because it was linked to his diet problems. Different words have varying degrees of sentimental attraction. An emotionally loaded word could be the name of a favorite food, a hated or loved person, a fulfilling interest, but also a less predictable word associated with an unusual experience etc.

The question is also whether it is not only taboo words that are processed involuntarily, but whether they can also be superhapaxes, which otherwise do not belong to common taboo words or swearwords. So far, only the processing of words representing emotion names or swearwords has been investigated, but not words outside these categories. Thus, we want to attempt to identify such words in authors and then elicit and measure responses in the laboratory based on the stimulation of these individuals with these emotionally laden words.

The above-mentioned research on the processing of words representing emotion names has been carried out, for example, by Bock (1986). Bock examined the conceptual processing of affectively positive, negative and neutral words, where the positive words were the most memorable ones, then the negative, and neutral words were the least memorable. It seems that "words are evaluated for their emotional content at an early stage of information processing as defined by Craik and Lockhart (1972)."

It is important to remember that there are emotional differences between people and people have different natures. Some people show less emotion and exhibit rather cool rationality, while others show stormy reactions to the smallest stimuli. The question is to what extent such two opposite poles are different neurologically. As an anecdote we can mention in relation to the lexicon a man who always said "operative" and was by nature always "sewing something with a hot needle."

In relation to emotion processing, we might also consider how broad a set of basic emotions to work with. It would be easiest to work with polarity features only, the question is how concise this is in terms of linguistic processing. In addition, current research claims that there are 27 basic emotions (Cowen & Keltner, 2017). So these are universal categories and in this context we also need to consider universal semantic categories (we can decompose lexical meanings into sublexical ones) and their representation along with emotions, for example through WordNets, but the crucial point is that they should be individualized, each person will have slightly differently structured meanings and hence their association with emotions. Although universal categories are not entirely well documented, linguistics needs them for its representations so far. There will be bruntal effect in these WordNets that must be captured appropriately.

Studies of the brain (cf. Olson, Plotzker & Ezzyat, 2007) show that the temporal lobe pole has an interesting role: it functions as a convergence zone where concepts – semantic memories (accumulated general knowledge, not episodic events from our lives) – are infused with

emotional and personal meaning. Individual concepts are also stored here, abstracted from perceptual representations.

Since we still do not know the reasons behind the size of the segment length of 6250 words, we can try pushing the threshold to a lower level, e.g. 5000 words, and compare the results. At this segment size, low-frequency units with a frequency greater than one in particular do not seem to be telling, though still worth exploring. The question is how the threshold relates to the size of an individual's memory. Is 5000 words per segment appropriate for some and 6500 for others? Does it have anything to do with the individual's active vocabulary? Older people with an eroded mental lexicon (they show increases in tip of the tongue phenomena) might have different requirements than children and young adults.

Another consideration, besides the question of the gradual decrease in segment size with age, is whether there is also a gradual increase in segment size when a child begins to learn to speak and increases his or her vocabulary as the child grows older.

It would also be interesting to ask the question of the hapaxedness of numbers. The processing of numbers may present extra cognitive load and possible pauses and hesitations. It would be interesting to see if people have any unconscious favorite numbers. As Faltýnek says, for Kerouac it was the number eight, we find thirteen thirteen times. It would also be worth looking at how the influence of translation obscures hapaxes when the translator uses translation crutches, and whether hapaxes translate into other languages spoken by the speaker.

Mental lexicon

Mental lexicon is not (yet) a textualizable object (or it might be only to some extent), which means that we are not able to extract the complete dictionary from it. What we can do is letting volunteers listen to spoken production or letting them read chosen texts, meanwhile we monitor their brain activity. Creating a procedure during which the subjects would come up themselves with the content of their lexical memory is also possible, but it does not provide any guaranty of listing all the word forms contained.

Maybe we do not really need complete mental lexicons of individuals and mere analysis of their monologues would be sufficient. Such monologues, however, must be unedited and unprepared, otherwise the subject would be primed to use pre-selected items. Spontaneous production of one individual person (when there is no other communication participant and ideally also no lexical stimulation present) is probably the best reflection of the organization of the mental lexicon we can obtain when it comes to spoken production. In dialogues, individuals are primed by the lexical choices of other conversation participants.

To begin with, a literature search was conducted, during which at least 64 notions (including a variety of mental lexicon models, lexical access models, mental lexicon hypotheses, and paradigms) relevant to our research were identified.[3] Many of these are unsatisfactory for us, but the list will nevertheless be useful for later interpretation of experimental data in the context of cognitive linguistics and neurolinguistics. It is also tempting to design our own model to take into account the new hapax evidence and ideally also the evidence we have from the literature we have surveyed.

[3] Their list is included in the appendices.

Considerations like how the brain constructs things come into play: how does it deposit things throughout life, what centers (if there are any) enable the brain to improvise with the stock. How are the individual constructs connected? What is stored in memory and how? Based on our current knowledge of neurobiology, one would expect network models of the mental lexicon to be the most relevant.

Evidence to consider

As Faltýnek (2021) suggests, authorial hapaxes are an indication that mental lexicon is a hierarchic network with frequency layers and supports De Deyne's model (2016). Hapax evidence weakens the notion of the independence of semantic memory as limited to purely factual information; it seems that we could store factual information along with episodic information, that our learning of facts might be influenced to a certain extent by episodic memories. As Faltýnek reflects, some items in the mental lexicon could be strong nodes that represent entries into certain areas of the lexicon. Thus, superhapaxes could be these nodes that allow entry into lower frequency word layers or topic neighborhoods. Within the paradigm of conceptual metaphor (Lakoff & Johnson, 1980), as Faltýnek writes, superhapaxes as conceptual metaphors are "more integrated in the author's conceptual framework." Now it would be interesting to explore how much of a metaphorical role superhapaxes have, how many meanings they can take on, but it seems more likely that they could be tied to one particular context, i.e. one (or a few other) metaphorical uses.

The study of memory

After examining 64 models of the mental lexicon, we find out that it is better to work with more general models, i.e. models of memory, because it is the functioning of memory that underlies the functioning of the mental lexicon. It seems that using memory models is a much better way to explain the existence of hapaxes. Let us first describe the functioning of verbal memory in a brief and clear way. This of course entails the risk of oversimplification, so the following lines should be taken only as a rough approximation.

The information we are processing at a given moment is located in working memory in the prefrontal cortex and to some extent in hippocampus. In order to remember a number of blocks of information that enable us to understand e.g. sentences, etc., they are stored in short-term memory. The short-term memory has a storage for 4-7 items (so-called magic numbers).[4,5] If we process some information more often, more times, it is stored (consolidated) from working memory into long-term memory, in the left temporal lobe.[6] When learning motor plans, this is more the responsibility of the thalamus.[7]

The hippocampus mediates (perhaps among other parts of the brain) the formation of new memories of experienced situations (stored in episodic and autobiographical memory).[8] It also receives and processes emotional data from the amygdala. Returning to the location of an emotional experience might resurrect that emotion.[9] "The right head of hippocampus is more

[4] For the number 7, see Miller (1956)
[5] For the number 4 and some reflections on Miller (1956), see Cowan (2001).
[6] For example, cf. Salvato et al. (2016)
[7] cf. for example Bosch-Bouju, Hyland & Parr-Brownlie (2013)
[8] Eichenbaum (1993)
[9] Gluck, Mercado & Myers (2014)

involved in executive functions and regulation during verbal memory recall. The tail of the left hippocampus tends to be closely related to the verbal memory capacity."[10]

Long-term memory has consolidation and reconsolidation mechanisms, both of which are chemically different:[11] when we first learn information, we store it, but after a while the brain does a clean-up and recycling so that old information can be re-evaluated. Thus, this concept is somewhat forgotten (e.g., we learn or internalize a grammatical rule and already use it, but suddenly forget it and make a mistake, which we realize through self-monitoring, for example, and this repetition reconsolidates the rule).

Baddeley's (1974) model describes that working memory includes a phonological loop, and a visuo-spatial sketchpad. Auditory input uses the phonological loop (neurophysiologically connecting Broca's area with other temporal areas)[12], which has a short-term (rapidly decaying) phonological memory and an articulatory repetition system that prevents the decay of certain cues. The visuospatial sketchpad processes visual information.

It is already generally known in linguistics, that emotionally loaded words such as taboo words and vulgarisms flow through the amygdala. The amygdala mediates hasty reactions to emotionally charged stimuli.

The second most influential memory model, called the search of associative memory model (Atkinson & Shiffrin 1968; Raajimakers & Shiffrin 1981), in addition introduces for us somewhat relevant concepts of autoassociation (self-association in long-term memory), heteroassociation (inter-item association) and contextual association (association between the item and its context).

Psychologists divide memory into explicit (or declarative) and implicit (procedural). Explicit is further divided into episodic, semantic and autobiographical.

Emotional arousal improves the memorability of information, i.e. enhances consolidation. Stimuli without emotional load and context are more prone to extinction compared to items without emotional load (LaBar & Phelps 1998).

Some approaches also describe a system called a mental syllabary from which motor articulation programs are retrieved (Cholin 2008, Brendel et al. 2011, Kröger & Cao 2015).

Phenomena

We know from the literature that we have to work with at least two key effects that manifest their powers in the mental lexicon: the frequency effect and the contiguity effect. To these two main ones we can add other relevant effects manifested in memory such as recency effect, primacy effect, serial position effect, phonological similarity effect, articulatory suppression effect.

The frequency effect is an effect manifested especially in lexical decision tasks, where subjects have to decide as quickly as possible whether the word or pseudoword presented to them is a word or not. For this, reaction times and error rates are measured. Words that occur more frequently are recognized faster.

[10] Vanchakova et al. (2008)
[11] Cf. Tronson & Taylor (2007)
[12] Unger et al. (2021)

However, there are a number of other strange phenomena associated with the frequency effect that are not well understood: it seems that recognizing very common words when learning a word list is harder than recognizing rare words – a take called the frequency mirror effect, cf. Glanzer & Adams (1985); Hulme et al. (2003); Park, Reder & Dickison (2005); Duncan (2013).

The contiguity effect is manifested by the fact that items presented simultaneously are more likely to be remembered together. If they appear together frequently, their contextual similarity is probably greater. Thus, in the brain, such ideas will be associated and when one is recalled, the recall of the other will be faster. A model for the contiguity effect was developed by Howard & Kahana (2002) and Sederberg, Howard & Kahana (2008): presented items activate the temporal context that was active at the time of the original encounter with the item (or, maybe more probably, the context that has the longest temporal occurrence together with the target word). The temporal context can then prime entire groups or lines of thought (the assumption of spreading activation – not to be confused with the Spreading Activation Model described below). The contiguity effect sounds very consonant with what Faltýnek jokingly calls the bruntal (or Bruntál) effect: Bruntál is considered an ugly city with beautiful surroundings, this is metaphorically applied to hapaxes and their context.

Figure 1: Visualization of the Bruntál effect. Author: Klára Faltýnková.

Recency effect is relevant for short-term memory and learning, but not for long-term memory. It means that the last few items will be remembered well. However, simple distractors can cause an item to be forgotten because it has not yet managed to form an association in long-term memory.[13]

[13] cf. Raajimakers & Shiffrin (1981)

Primacy effect shows that the first few items have a better chance of being remembered. The first item has stronger autoassociation, heteroassociation, and contextual association, leading to greater associative strength.[14]

Serial position effect combines the two previously mentioned phenomena, i.e. fresher memories are remembered better, and what is also remembered well, are the first items mentioned. Howard & Kahana (2002) explain this via the temporal similarity and stochastic drift. The point is that the items that are to be remembered have a certain context in which they are learned and this context is getting continuously forgotten and so the overall similarity is getting lower.

The existence of phonological similarity effect somewhat increases the importance of phonological encoding of language and alerts us to the fact that language could be encoded primarily phonologically. However, it does not represent much change for our research. Already Baddeley (1966), the author of the working memory model, assumed that short-term memory is dependent on auditory encoding (as opposed to long-term memory). Words that are phonologically similar are harder to remember than words that are dissimilar. On the other hand, long-term memory relies more on semantic relations.

Articulatory suppression (people are asked to say something irrelevant) effect provides evidence that short-term memory span is limited (cf. Baddeley 1975).

Associations

A major finding in the literature was the S-P (syntagmatic-paradigmatic) shift or S-P-S (syntagmatic-paradigmatic-syntagmatic) shift described in Petrey (1977), who makes an interpretation of association data collected by Entwisle (1966). The point is that young children show a tendency for syntagmatic associations, whereas adults show a tendency for paradigmatic associations. Syntagmatic associations mean associations with a word next to the stimulus word, e.g. to attribute a noun, Petrey lists the subject "flour" to the verb "add," etc. Paradigmatic associations mean vertically selecting from the possibilities of words that can come to the same position in a sentence. E.g. for the word chair it could be table, instead of "black" it could be "white." But the crucial thing is that in some cases (for some word associations) adults revert to a tendency towards syntagmatic associations, so it would be possible to say that their context "solidifies": the word "hermetically" is tightly associated with the word "sealed," the word "gallop" is semantically restricted to "horse."

It is this solidification of associations that might explain the hapax phenomena, where hapaxes tend to occur in the same context. Moreover, some models of memory suggest that we learn all information along with all context: situational, emotional, etc. In particular, we might have stored low-frequency words right along with the contexts that were around that word when we first encountered it (or during some important encounter, of which there were not many in total).

According to Petrey, this shift occurs in relation to the separation of episodic and semantic memory. Endel Tulving (1972), one of the most cited psychologists, distinguished between episodic and semantic memory on the basis of psychological experiments. We have the ability to consciously recall previous experiences and events, which episodic memory allows us to do, but we also have an organized store of general knowledge, which semantic memory allows us

[14] cf. Atkinson & Shiffrin (1968)

to do. As Petrey notes, subjects who had neither syntactic nor semantic control over a stimulus recalled the circumstances during which they perceived it.

This would mean that we have word networks stored in our brains that have individual structure according to the contexts in which we encounter words. The division between semantic and episodic memory would not be so strict. Now the data to be processed are in the form of groups that do not follow this principle, the work is to compare the groups of hapaxes that have a solidified context and the hapaxes that do not demonstrate significant tendencies to be attracted by certain contexts.

In other words, speakers recover a partial preference for expression relations over their overall preference for relations at the semantic level, where they stabilize certain ways of expressing themselves. People behave more predictably with emerging adulthood due to the increasing of paradigmatic versus content relations, while at the same time there is a semantic restraining that is individual and enables us to identify the author.

Recall

The recollection of words is localized in Broca's area (Thompson-Schill et al. 1997). Sometimes it happens that we cannot remember a word. Although we know it and it should be stored in the mental lexicon, we probably use it less frequently and accessing it is difficult because of low myelination. We are often unable to recall the initial syllable, which would support Forster's model of autonomous retrieval as well as the cohort model: "What is the name of that yellow flower?" – "D-... d-... daffodil!" We have a sense of what a given word should sound like, so it is possible that words are not arranged only by onset. It is possible that they are mentally arranged according to other structural features as well. Sometimes the feeling is wrong. This phenomenon, the tip of the tongue phenomenon, resembles parapraxis.

There is free recall (without cues) and cued recall (when cues are available), or even serial recall (which has been studied more in connection with short-term memory, while long-term memory would be relevant for research on remembered poetry and reciting individuals). Recall shows many word associations, as we read in Petrey (1977), the assumption that free association activates only semantic memory does not hold (while pointing out the opposition of syntagmatic and paradigmatic); it is apparently also dependent on episodic memory: subjects with neither syntactic nor semantic control of the stimulus recalled the circumstances where they had perceived it."

Stille et al. (2020) nicely describe speech processing based on psychologically delineated memory types. First, we need to distinguish production and perception, which have different pathways. For both pathways, data are retrieved from a mental lexicon in long-term memory into working memory (here they cite Vitevitch et al. 2012). "Whether information is stored depends on several factors such as attention and the importance of the information."

Stille et al. continue: "During speech production, concepts are activated for a planned utterance, and associated lemmas and phonological forms are subsequently activated and then retrieved from different levels of the mental lexicon." Further, they state that "associations at the concept level are based on semantic similarity. Associations occur not only with respect to categories (like "animal" or "object in a room") but also with respect to more specific attributes such as size, shape, and color (McGregor and Waxman, 1998). Associations can be different from subject to subject depending on differences in personal experience both during and after speech acquisition. In general, associations are built up within the concept level on the basis of features like "has four legs," which establish similarity relations between concepts like "dog" and "cat"

as well as between the concept level and the word or lemma level, since each lexical entry is directly linked to one or more concepts (Lucas. 2000)."

Drawing from Levelt et al. (1999), Stille et al. (2020) also mention that phonologically similar lexemes share more associations – it depends on the level of similarity. Later, based on Meteyard & Bose (2018), they introduce the hypothesis "that phonological cues support the retrieval of phonological information for a target word, while semantic cues support the retrieval of semantic information for a target word."

Priming

It would seem that hapax evidence would shake up what we know about priming. However, if we take the findings from psychology and add our topic/context priming (which of course has yet to be experimentally verified), it should fit well into the existing picture. Priming can be lexical, semantic, phonetic and orthographic, morphological, syntactic. Among the semantic, we can distinguish in more detail translational (cross-linguistic) – this type should be investigated in the experiment proposed below in the chapter Experiment design suggestion, conceptual, cultural; associative or contextual priming is also known from the literature (the latter is not necessarily semantic, just some things occur together independently of semantics). Mediated priming also works, i.e. exposure to words that are one step further away: a neurological study by Sass et al. (2009) give the example of stripes for the word lion, when stripes have tiger versus the word chair. There exists even somewhat mysteriously sounding subliminal priming (cf. Van den Bussche et al., 2009), where the priming information is presented unconsciously.

About the models in general

As Stille et al. (2020) write, it is a shared feature of many models that they are realized as a three-level neural network, with mentioning the following list of authors: "Collins and Loftus, 1975; Garrett, 1980; Stemberger, 1985; Dell, 1986; Butterworth, 1989; Levelt, 1989; Caramazza, 1997; Dell et al., 1997; Levelt et al., 1999; Indefrey a Levelt, 2004; Indefrey, 2011." The individual levels have different kinds of information about individual words. Some models introduce conceptual networks (in some way, this is analogical to the approach of Fodor's mentalese) working above the mental lexicon and organizing it. The evidence for the existence of mentalese are the accounts that show the independence of higher order thinking from access to meaning, such as in Seidenberg et. al. (1982).

Semantic models

Hierarchical network model

One of the first proposed models of the mental lexicon was the hierarchical network model (Collins & Quillian 1969). According to it, word forms and meanings share lexical entries in the brain and a pyramidal arrangement of concepts is assumed, where its levels correspond to abstractness. However, such an arrangement has not been demonstrated because experimental evidence shows that access to more general concepts is slower than access to even more general concepts, and thus the thought process does not flow in this way along these assumed levels.

Spreading-Activation Model

The Spreading-Activation Model, developed by Collins & Loftus in 1975, suggests a non-hierarchical organization where each node of the network can be connected to any other node. A node can be represented not only by a lexical item, but also a semantic feature. Every individual person has a different organization of the items (which we strongly agree with in light of the hapax evidence).

As seen through the lens of present-day's evidence, this model contains a lot of serious flaws and can be labeled outdated. One of the main flaws is the fact that it does not take into consideration some important phonological, morphological and syntactic aspects of language processing that we have evidence for today, although it is based on evidence from some experiments with these language aspects that were available at the time.

A 1994 extension (by Bock and Levelt) supplemented the model with separated layers for forms and lemmata (thus making it more morphology-based). Every lemma should have its own lexical "pointer," pointing to the memory locus of the corresponding word form. Today, this suggestion seems very arbitrary and improbable because of the frequency effect.

Bock and Levelt (1994) drew mainly from the evidence with production errors. They cite semantic substitution errors ("She was handing him some cauliflower" instead of "She was handing him some broccoli"), errors of function assignment ("He was handing her some broccoli"), stranding ("You ordered up ending some fish dish" instead of "You ended up ordering some fish dish") and shift ("She was hand himming some broccoli").[15] These kinds of errors point to a tendency of the speaker to form utterances based on morphological rules rather than segmenting the speech into phonological pieces.[16] These errors might be caused rather by the motoric apparatus than the way of organization of the mental lexicon. Or the thought process might be too fast that the motoric apparatus was not able to adapt to it at the moment.

Other evidence used to support the Collins & Loftus model was based on experiments with sentence verification, which seems too abstract and too high (it is relevant for higher-order semantic processing, but does not seem to reflect the organization of the mental lexicon), cf. Holoyak & Glass (1975); Rips, Shoben & Smith (1973).

Adaptive character of thought model

The adaptive nature of mind model (Anderson 1996) posits a separation of meaning and word on the basis that it is possible to have a concept without a word but not a word without a concept. The model assumes the cooperation of declarative knowledge ("that") and procedural knowledge ("how"), which sounds good for our hapax evidence. According to it, complex cognition actually originates from this interaction. The network learns how likely a given word is to occur together with others, forming clusters based on functional context; we cannot argue against this with our evidence.

Procedural knowledge is stored in units called production rules, declarative knowledge in units called chunks. The chunks are formed by simple encodings of transformations in the environment (production rules). From a large database of units, context-appropriate units are selected by the activation processes behind the statistical structure of the environment. Whether it is the selection of a memory to recall, the categorization of objects, or the selection of a strategy, humans are sensitive to prior information and information about suitability for the situation. Although it is not a conscious process, people mix this information in a Bayesian optimal way.

This is, however, where the appeal of the model is slightly diminished. Individuals have individual tendencies to express themselves using the same words over and over again in given contexts. First, the statistical structure of the environment, while it affects the context-

[15] Bock and Levelt 1994: 947-949
[16] Bock and Levelt 1994: 976

appropriate selection of units, does not guarantee the selection of the most appropriate unit. Each person has his or her own solidified structure of words to use in specific contexts, which is different from that of other people. People are definitely sensitive to prior information and information about appropriateness for a given situation; the model is right about this, but it does not account for individual structuring and word use.

WordNets

WordNet databases contain lists of lexical entries for individual concepts or contexts (synsets). The defining feature of a synset is always a single concept. The synsets are connected by a hierarchical network that uses hyponymy and hyperonymy – we have somewhat criticized above such a tenacious adherence to hierarchicality. Humans do not have words and concepts in their mental lexicon structured exactly according to unbiased semantic facts, and neither can even scientists have coincident knowledge and opinions about the classification of units. A well-known problem with WordNets is that they cannot be used to predict semantic priming. So, for example, someone will consider a particular species of bird to be a member of one taxonomic family, while another will consider it a member of another family. Another person will not be able to distinguish the family or even the species and will simply call the bird a bird.

The structural semantics- or semasiology-based alternatives to WordNets could be the methodological clustering of Hallig and von Wartburg (1952), notional fields of Trier (1973), lexical arrays of Matoré (which constitute configurations of associations, keywords are in the center of thematic fields), cf. for example Matoré (1953). However, even these approaches are somewhat non-individual and would need to be individualized to make them useful for the description of an individual's mental lexicon. Without individualization, they do not represent sufficiently useful tools for us.

(n)ROUSE

(n)ROUSE, where potential n means "neural," is an abbreviation of responding optimally with unknown sources of evidence. The model ROUSE was introduced in 2001 by Huber, Shiffrin, Lyle and Ruys, in 2003 they published the nROUSE model (Huber & O'Reilly). The primary aim was to explain unknown sources of priming using source confusion and statistics. They claim that there can be random similarities causing priming or random unidentifiable sources of word activation (Shiffrin & Huber 2001). These claims are for us to be examined through the research of context priming and context learning together with vocabulary.

Models of lexical access

We have explored at least eight models of lexical access in the literature: logogen model, autonomous search model, cohort model, TRACE, NAM, PARSYN, Shortlist and Kawamoto's model with distributed representations. Due to the limited scope of this paper, we omit the analysis of lesser known models, i.e. NAM, PARSYN and Shortlist and Kawamoto's model.

There are many other factors that influence lexical access: we know that at least sentence context (Morris 1992), meaning frequency (Griffin 1999) and imageability (de Groot 1989, whose findings suggest conceptual thinking might not be based entirely on linguistic logic) play a role. When a word is ambiguous, its realizations within different modalities (phonological, orthographic, individual meanings) vary. Without the biasing sentence context, the dominant meanings are strongly activated, while secondary meanings are activated later, Simpson (1981) measured a 120 ms delay.

There are accounts that lexical access might be controlled differently in different modularities: Gollan et al. (2011) found that "Frequency effects were larger in production than in reading without constraining context but larger in reading than in production with constraining context." They call this the Frequency-Lag hypothesis.

The findings of one meta-analysis (Lucas 1999) on focus are also relevant. Focus may also influence lexical access. The lexical decision immediately after the multisyllabic word was faster than when the multisyllabic word was in sentence focus. Attention (which is most likely mediated by the thalamus, as the literature suggests, cf. Radanovic et al. 2003) did not affect lexical access content, although there is a trend toward a larger contextual effect during focus. Attention significantly affects the speed of the process. The universal reduction in the size of the weights in the network simulates processing outside the focus of attenuation. Smaller weights slow down and weaken processing. As the author of the meta-analysis points out, it is difficult to test for any small effects, but if we find consistently nonsignificant trends toward a significant difference, we should not ignore it. This is primarily an effect of sentence context on the activation of ambiguous word meanings. Almost every experiment testing the independent approach hypothesis found at least a trend toward biased activation. Individually, this trend means nothing, but overall it constitutes a strong affirmation of biased activation. This places sentence context among a set of factors: frequency, rate of presentation, attentiveness, all of which affect lexical access. The results support interactive activation instead of modular activation.

Some scholars (Luce & Pisoni 1998) propose "that the well-known word frequency effect may be a function of neighborhood frequency and similarity, and not a simple direct function of the number of times the stimulus word has been encountered." They point out that there could be relationships between the sound patterns of words in memory. To this one could argue that it could be a bias of people who are more poetic in their thinking. Just as only people who have learned syllabication have access to the syllables of words versus people who have not learned syllabication, what if some people who read poetry (or even create it) had a more structured vocabulary according to the similarities of words to rhymes?

Cohort model

The cohort model has an old and a newer version (Marslen-Wilson et al. 1978, 1987). It proposes a three-phase lexical access and assumes that simultaneous activation occurs for phonological neighbors of a word, while it does not occur for semantic neighbors. In the first phase, the initial syllables activate all words with a similar sound – and these words are the cohort. In the second phase, selection – narrowing of the cohort – takes place (discrimination of the best match, discarding of competitors, processes of activation and selection or recognition and competence take place, according to the second version of the model they continue until the so-called "recognition point"). And in the third stage, the integration of one particular item occurs (mapping syntactic and semantic information at the lexical level to higher levels of processing).

From the beginning, the model has been adapted to allow the role of context to exclude competitors, while activation is an acoustic approximation because it allows for certain coarticulatory changes (see Packard 2000: 288).

The cohort model has been discredited by many authors: Slowiaczek, Nusbaum & Pisoni (1987), Altmann (1997), van Heuven, Dijkstra & Grainger (1998), and also the TRACE model

by McClelland & Elman (1985). We must also criticize it, since our evidence does not support it: rather than cohort, we see semantic relations and conceptually uncontrolled and non-normative vocabulary and neighborhood effects.

TRACE

TRACE by McClelland & Elman (1985), which outperforms cohort in that it can explain underspecification or erroneous pronunciation of the word beginnings, has a perceptual processing mechanism and a working memory, and consists of three levels: auditory features, phonemes, words (the mind goes through each level and decides which word was heard). The mind uses auditory features, phonemic and semantic information to connect with what is heard. Higher and more abstract levels of knowledge can interact with lower levels of processing.

TRACE solves the problem of segmentation and the Ganong effect (Ganong, 1980), which are more phonological phenomena, and we mainly need the semantic aspect of things. However, TRACE also involves semantics to some extent: "the semantic and syntactic context further constrain the possible words which might occur" (McClelland & Elman 1986). In this respect, TRACE appears to be more attractive.

Logogen model

The Logogen model (Morton 1969) focuses on explaining the frequency effect, but otherwise it is poor and it can be judged uninteresting.

Autonomous search model

Autonomous search model (Forster 1976) uses libraries with catalogues, where the catalogues represent orthographic, phonological and semantic-syntactic aspects of the language and only one catalogue can be accessed at a time. First the initial part of the words is searched and then only the exact location of the word is searched. So the first stage is finding the catalogue and the second is finding the book. The catalogues rank the words according to the frequency of their use and the search proceeds until a perfect match is achieved. If one is exposed to a non-existent word, it takes longer to reject it because a search of the entire catalogue is required.

Connectionism

Connectionists try to move away from biological reality and base models on neural networks. There are many connectionist approaches, one of the first being the influential interactive activation model of McClelland and Rumelahart (1981). They range from multilevel distributed systems to models emphasizing Hebbian learning, Fodor's language of mind, or Smolensky's integrated connectionist-symbolic cognitive architecture (cf. Smolensky, 1990; Legendre, Miyata & Smolensky, 1990). In this paper I will touch on them only lightly, leaving a more detailed analysis (which is certainly tempting) for later works.

The disadvantage of most interactive activation models is the inability to learn, unlike distributed processing models. Thus, distributionalist models would definitely be more complete and appropriate for us.

In the neural networks of connectionist models, the neural nodes typically represent individual words and the words are activated through certain activation patterns. The nodes can be also sublexical entities.

To evaluate individual connectionist models, research on how they address frequency effects and contextual priming is necessary. They must be individualizable and allow the integration of superhapaxes that have regular occurrence after some time.

We can also consider the design of our own connectionist model, assuming individual structuring of the lexicon and emotional, topical and contextual links to words, counting with the S-P-S shift.

Multilingual models

BIA, BIA+

The first version of the model, BIA (bilingual interactive activation) was developed by Dijkstra & van Heuven (1998) as a connectionist model of bilingual word recognition model. It is language non-selective, meaning that it automatically activates a word in both languages simultaneously. According to Dijkstra, van Heuven and Grainger (1998), language-selective hypothesis enjoys only little empirical support. The model contains four levels of representations: letter features, letters, words, language labels (or nodes). In this sense, it is best adapted for reading. Language nodes may inhibit word candidates from other languages, the best candidate becomes the most activated one. We do not yet have hapax data to evaluate such associations, but they are forthcoming. The most natural assumption would be that it will depend on the way the word is learned and its subsequent use in context: if the word was used in context with a word equivalent to it in another language, then there is likely to be faster access between the two, the question is whether analogization in the mind alone is sufficient. In this respect, both hypotheses (linguistically selective and non-selective) would appear to be flawed and irrelevant.

The improved successor model, called BIA+ (Dijkstra & van Heuven 2002), newly incorporates phonological and semantic lexical representations and specifies a purely bottom-up nature of language processing, while our hapax evidence points to the opposite, top-down construction. Both words and phonological representations then activate semantic representations and linguistic nodes that indicate affiliation to a particular language. All this information is used in the decision subsystem to complete the task. It works with two types of context: linguistic (processed by the word identification system) and non-linguistic (processed by the task/decision system).

If we open Petrey (1977) again, we come across the remark that these two contexts could correspond to different types of episodic information: the verbal context, or the co-occurrence of a word with other words, and secondly the situational context of the word, i.e. the situation in which the word is used. Petrey here goes on to criticize the over-separation of the functions of the different types of memory (autobiographical, semantic) and points out that the evidence of word associations suggests that retrieval is dependent on episodic memory.

The parallel approach of the BIA+ model assumes that language is nonselective and that both potential word choices (a word from the native language and a word from an L2) are activated in the bilingual brain when exposed to a stimulus. The study (Dijkstra & van Heuven 2002) shows semantic priming for both languages: "the N400s for interlingual homographs were sensitive to the relative frequency of their readings in both the target and non-target language." Furthermore, an individual cannot consciously focus attention on just one language, even if they try to ignore it (it would be interesting, however, to look at hyperpolyglots using Krashen's method and thinking in the target language). The nonselective approach appears to work for semantic, orthographic, and phonological activation. The L2 temporal delay assumption is based on the frequency of word use: high frequency words correlate with high resting activation potential and vice versa.

Network models? Or other models?

In general

Theoretically speaking, based on what we know from biology, mental lexicon should be realized as a neural network that stores information. Individual memories are stored in neurons. To enable the flow of thoughts, attention spreads between neurons via synapses, each neuron has thousands of synapses (i.e. can be connected to thousands of other neurons). With the human brain containing around 100 billion neurons, we arrive to the number of possible synapses as high as 1 quadrillion (Zhang, 2019). During the learning process, we strengthen the more frequently used synapses. Synapses from stored words could lead to different types of information such as emotion, context, situations where the word is used, spaces (physical and social) where the word is used, etc. Some neurons may represent more important nodes of the network through which more traffic passes. Since such a complex network cannot be reproduced at least yet, all models will necessarily be mere over-generalizations. Our aim, therefore, is to find the most accurate generalization possible, or to be able to create an individual model of one person's mental lexicon; this would be psychological profiling.

Multiplex model with "explosive learning"

One of the most recent models (Stella et al., 2018) emphasizes the importance of explosive learning, i.e. a type of learning that remains undefined within the study, but can be described as involving a sudden acquisition of an important item (or rather items) that will connect many nodes of the mental network. Such items are highly polysemic. The authors state that the most important event of explosive learning occurs around age 7. In their simulation, this newly emerged cluster involves 1173 words, which account for 13.8% of the lexical items.

It contains four layers that code for different relations among words: free associations, synonymic relations, taxonomic relations and phonological similarities. The datasets for the simulation come from various independent sources. For free association (where free association means "A reminds one of B"), the EAT, or Edinburgh Associative Thesaurus (Coltheart 1981: LRC Psycholinguistic Database) was used.

The EAT (see the description in Pajek Datasets, 2003) is a dataset of word associations that were collected from the study subjects: the researchers presented a word to study subjects and they say the first word that comes to mind, the association is not semantically labelled (as a synonym, antonym or otherwise), it is merely an association; norms were collected by expanding the network from a core set of words. The core set was based on the 200 stimuli from Palermo & Jenkins (1964), the 1000 most frequent words from Thorndike & Lorge (1944), and the basic Ogden English vocabulary (1954), responses were collected for these words, then the responses were used as stimuli, etc. Data collection ended at 8400 stimulus words. Each stimulus was presented to 100 people, each receiving 100 words. This gave rise to 55732 network nodes in the EAT. Subjects were mostly students from different universities aged 17 to 22, 36% male. The data was collected between 1968 and 1971. The database has two sets: stimulus-response and response-stimulus. The stimuli were presented to the subjects printed in random order (to minimize priming), a different order for each subject, the subjects had to write the words, they had to do it as fast as possible, they completed it in 5 to 10 minutes.

For taxonomic relations ("A is a type of B") and synonymy relations ("A also means B"), the authors of the Multiplex model used the WordData from Wolfram Research (WordData source

information),[17] which mostly overlaps with WordNet 3.0[18]. The dataset used for phonological similarities was from the same authors (Stella & Brede 2015), based on WordNet 3.0. All layers are undirected and unweighted. Words in the multiplex representation must be connected on at least one layer.

Free associations are similarities in semantic memory, e.g., when people respond to the cue "house" with words that remind them of home ("bed" or "home"). Free association networks play a prominent role in the capture of word acquisition in toddlers (Hills et al. 2009, Stella, Beckage & Brede 2017) and in word identification (De Deyne et al. 2016, Collins & Loftus 1975). The authors refer to studies pointing out that synonym networks also play a role in lexical processing and that the hierarchy provided by taxonomic relations deeply influences word learning and processing. Phonological networks provide insights into the competence of similar-sounding words for interchangeability in word identification tasks.

For the linguistic attributes, different sources were combined in the multiplex model: word frequency from OpenSubtitles (Barbaresi, 2014), a dataset of movie subtitles whose frequencies are considered superior to those in classical sources with respect to explaining variance in reaction time analysis from lexical decision experiments (Keuleers et al. 2012, Brysbaert, Warriner & Kuperman 2014). Concreteness scores (Brysbaert, Warriner & Kuperman 2014) and age of acquisition were taken from Amazon Turk experiments, allowing for large-scale data collection and confirmation of previous findings based on small-scale experiments (Kuperman, Stadthagen-Gonzalez & Brysbaert 2012, Brysbaert, Warriner & Kuperman 2014). Polysemy was quantified as the number of different word definitions in WordData from Wolfram Research, overlapping with WordNet. Response times were taken from the British Lexicon Project (Keuleers et al. 2012) and indicate the response time in ms for identifying individual words compared to non-words.

For research on the formation of superhapaxes and their establishment in the mental lexicon, it would be interesting to look at the development of speech from childhood onwards and observe people's speech over a long period of time to see what the role of words that would later become superhapaxes would be. What specific words children learned and when, to compare general and individual accounts. A database is available from Entwisle's work Word Associations in Young Children (1966), which could also provide interesting insights. There is also a Childes database, but the records are too short. It would be worth finding out at what point a more or less steady network of about 3500 superhapaxes develops. In the beginning, children learn their name, which in terms of frequency, may behave like a superhapax. Although children learn the more common words first, they will certainly learn some specific hapaxes as well.

Multiplex was one of the null models against which the authors compared normative acquisition, i.e., they modeled and ascertained from acquisition types when LVC (largest viable cluster) would develop. The question for us is, did its development occur because of the acquisition of specific synapses or specific words?

Stella et al. (2018) write that no difference was found for either feature, which according to them suggests that LVC arises due to higher-order synaptic correlations rather than local topological features of any degree or psycholinguistic attributes. Thus, it is the global

[17] Wolfram (2017)
[18] Miller (1995)

distribution of connections that drives the explosive development of LVC. Connections important for LVC formation may be acquired earlier, but LVCs emerge later, after some key pathways are added.

The authors also attempted to rearrange the connection while preserving word degrees, which resulted in layers that exhibit non-trivial LVC; when randomly shuffling some of the connections, there is still a tendency for LVC integrity (especially when free associations or errors in phonological transcription were cancelled out); by using the null hypothesis of shuffling words independently of any layer, they ensured that inter-layer correlations are not preserved, but the network topology remains the same (but there are no LVCs).

So, in the future we could work with this interesting model, and include the connectivity and influence of cognitive objects like emotions etc.

Another point is that we do not have access to layers; they are distinguishable by revelation, made available through learning (like syllables, orthography, while phonology and semantics are learned intuitively without identification – except by the linguists); but in reality they do not represent real pathways, but a tangled mess that cannot be untangled. How then does the higher semantic plan and conceptual thinking take advantage of this cannot be answered, it depends on the nature of consciousness and thought, and in the context of this paper we are only interested in the information stored and accessed – but in the sense of the movement of the stream of consciousness, not the nature of the access itself (the principle); knowledge of all subject areas connects concepts and therefore the vocabulary.

A new model suggestion for the hapax evidence

Many of the mental lexicon models were not based on psychological evidence, but only on linguistic experiments, which in many cases dealt with matters that may not be so much related to the organization of the mental lexicon. In comparison with this, Eleanor Rosch's prototype theory (cf. 1973, 1975a, 1975b, 1976, 1981), which is also somewhat influential in cognitive linguistics, and is cited by Lakoff (1990), seems to be attractive.

So let us assume that everything (but we will not deal with motor programs for the moment, which we will now omit from our model because they are not relevant) is stored in the brain in the form of an approximation to the prototype.

While many models try to explain morphological processing through hard-to-prove computational operations, etc., the simple explanation seems to be that the brain approximates (sometimes perhaps via a lemma or via the most frequent form) and can evaluate the morphologically most appropriate outcome based on some prototypical ideas about the system and analogies. These prototypical ideas are unconscious. Prototype theory has already been applied to morphology in linguistics, e.g., by Bednaříková (2011). We simply associate a word with a prototype, yet we also have prototypes of morphemes, which allows us to make an approximation based on context and the underlying sentence pattern. The aspects of a word for which an approximation needs to be made include the phonological component, the orthographic component, the referential and semantic component, the suitability of use in a given register, possibly the syntactic function or even collocational occurrence; morphological information may also be added, especially for irregular words.

From a syntactic point of view, we may have certain prototypical ideas of what a sentence looks like, this may be, for example, the base sentence formula, which together with other word forms

we encounter, gives us a certain prototypical idea of grammatical phenomena, and we are thus able to determine the correct form intuitively in a given context without deep thought about the language.

Prototypes would explain one of the basic principles of the human brain: efficiency. Unlike models that propose that the brain could store everything (and create new neural memory tracks for everything), the hypothesis of approximating to a prototype sounds much more plausible. Of course, the prototype model needs to be significantly strengthened. For example, a wug-test could help.

It would be (because of the biological background) a connectionist model of lexical memory with layers or core and periphery (core is active vocabulary, periphery is passive), there are transitions between them, the active vocabulary drive can reach into the passive vocabulary. It is necessary to think about how this drawing relates to hapaxes. Naturally, the question also arises of the connection between the size of the active vocabulary and the number given in the Multiplex model: 7000, this is close to 6250 and could be related to explosive learning. The question is to what extent superhapaxes are connective; we would need to measure their importance in the texts and have more data regarding their topic priming.

Transition between active and passive vocabulary

There are ideas that active vocabulary is limited despite a large passive vocabulary. It is nicely described in an opinion article (The k2p blog, 2017) titled There is a cognitive limit (the Wordsmith number) to the number of words you can know?: "It cannot be memory capacity in the brain that sets the limit. My hypothesis is that just like there seems to be a cognitive limit to the number of significant social connections a person can maintain (the Dunbar number – averaging around 150 with a minimum of around 50 and a maximum of perhaps 250), there is a cognitive limit (the Wordsmith Number) to the size of the active vocabulary that a person can maintain."

The article continues: "Even for those who are multilingual, the sum of the words they command in all languages seems to be limited to be no different to those who are monolingual." The author even poses a hypothesis: "My hypothesis is that there is a stable level – the Wordsmith Number – which the brain establishes. It is a cognitive limit to the size of the active vocabulary that a person can maintain. It is established by the manner in which the brain learns, stores and retrieves active and passive words. It is a dynamic level and varies as our activities change (reading, writing, speaking, diversity of social relationships ..). Words that are not active are shunted out of active memory. In very rare circumstances is a Wordsmith Number of greater than about 30,000 established."

Although this is not a scholarly article, its ideas are surprisingly appealing for our research and in any case would be worth verifying. The hypothesis put forward by the article would explain some hapax phenomena and why we prefer certain words in certain contexts.

Could there be a physiological limit to active vocabulary? That is, there would be no more neural layers, and we would have to reconfigure the active network and set up new connections that we use, dragging more frequently used words from the passive vocabulary into the active vocabulary. Some of the firings will be in the nature of tighter nodes that will embody pathways to the passive vocabulary. The reason why active vocabulary should be limited is because it would be uneconomical for the brain to maintain unused connections. As nicely analogized by Faltýnek, we can imagine that if we bridge Bruntál across the highway, there will be less driving

to Opava, then it will be demolished and there will be more driving south, i.e., a different road along the main ones to the passive vocabulary.

Interpreters who use a lot of words may appear as a counter-argument. However, they have had intensive training and it is possible that this ability can be trained, i.e. the size of the active vocabulary can be increased.

So let us assume that the empirical limit of 6250 words is related to the size of the active vocabulary. Could it be just the size of the active vocabulary? Rather not, because different individuals (cf. people with little education and interpreters) have different sizes of active vocabularies. Less educated people who read little do not use a very large vocabulary, while practiced speakers and polyglots are no doubt capable of using many more words than just 6250. Pushing the limit higher does not seem possible on the basis of Faltýnek and Matlach's experiments. It is still necessary to study the hapaxes of children and old people (since according they are subject to annual in vocabulary because of age, TOTs are more common when trying to access low-frequency items[19]), to create a possible fitness scale for lexical memory if needed, but otherwise all we can assume is that the 6250-word threshold is already reliably differentiating enough to distinguish one person from another.

To determine active vocabulary, Productive Vocabulary Levels Tests, among others, are used to measure knowledge of word frequency bands, e.g. Laufer (1998) uses the 2000, 3000, 5000 and 10000 bands. In addition, Laufer distinguishes between passive, controlled active and free active vocabulary: "The distinction between controlled and free active vocabulary is necessary as not all learners who use infrequent vocabulary when forced to do so will also use it when left to their own selection of words."[20]

Now there is also the question of the degree of knowledge of the word, one does not have to know all the uses. However, on the contrary, what is essential for us is that someone uses a word specifically in his individual deep-rooted context. What is relevant, of course, given that a superhapax can be any form of a word, is the definition of the word in the mental dictionary – it might be the case that if a word form is frequent we store it in order to access it more quickly, whereas less frequent forms derive presumably from the base form (it may be lemma, although this is not certain).

Vocabulary size measures can give very different results based on the word definition used: apart from word form, it can be lemma or even a word family, which would decrease the number significantly.[21] The level of activity of a word in the vocabulary could be different at different levels if a person encounters a word more often by hearing it or seeing it written, versus using it themselves in spoken or written form.

6250 is a construct that, while having some relation to the active vocabulary, is still abstract, non-physical in nature. It's such a large enough number of words that it simply takes the author back to his favorite topics, and the number can be reduced because one is focused on certain topics. It simply depends on the nature of the samples chosen.

Experiment design suggestion

To test the hypothesis of learning words together with contextual information, a clinical study could proceed as follows: in the first part of the experiment, subjects would learn a list of chosen

[19] Brown 1991
[20] Laufer 1998: 257
[21] Brysbaert et al. 2016

foreign language words. The sample of words would be selected in a suitably representative way, i.e., say at least 20 words from three different semantic groups (i.e., 60 words in total), and we would have at least 20 subjects learning the same words. The words would be taught by different methods: a part of the word list (or possibly a group of the subjects – or a combination of both approaches) would be taught by Krashen's input method,[22] i.e. the words would be explained through the target language without using the mother tongue, while the other word/subject group would be taught through translation into the mother tongue. After a few weeks, we would prime the subject and measure their reaction times according to which language and method they were primed through. The assumption would be that presenting an equivalent word in the native language when testing a word learned by Krashen's method, i.e. only through the target language without a stronger connection through the native language, would cause a longer reaction time because the person had not made that connection. In addition, words learned through translation should have a stronger connection between languages and therefore faster access and reaction time.

We can assume that translators and interpreters make connections precisely by constantly thinking about equivalences between languages. A person who speaks more than one language but does not translate will not have such strong inter-language connections, although it cannot be said that he has none, because he or she will be somehow thinking conceptually and translating something into his or her mother tongue, for example, he or she may come to connect equivalents between languages in some less intensive way.

Another experiment could investigate contextual word learning: people would again learn some words and then reaction times would be tested when priming the context in which they learned these words. If they were primed for a context that was semantically unrelated (i.e., not semantically typical, or default for the word) in which they learned the word, the hypothesis would be that they would still have faster reaction times when priming that context. The second group would be priming the semantically typical context, and the control group would be priming a different, unrelated, unlearned context. The situational context can be as simple as eating an apple. We need to see if indeed the data will suggest that the general assumptions are valid, but defined by the individual structuring of the mental lexicon.

The experiment could also be clinical and involve investigating neural processing, but it would need to be carefully designed and give clues as to what variables to observe based on e.g. existing neurostudies of taboo words processing etc.

Methodology

Our work consists of doing analyses on both written texts and transcribed monologues. Faltýnek & Matlach (2021) examined the hapax structure and found that low-frequency items are an important mean of cohesion. Low-frequency items constitute around 40% of words in texts, which is why we think they must have many important functions and that they are useful for forensic applications. People have a high probability of repeating a particular word with a similar dispersion and constant frequency in their production.

[22] For Krashen's methods and hypotheses, see his work from 1970s and 1980, i.e. 1977, 1982, 1989.

"Authorial hapaxes could represent network nodes of certain lexicon frequency layers through which nodes the speaker more often enters certain areas of the network – this also determines the structure position of the rest of the lexicon in the area."[23]

We project the hapaxes back into the text, looking at how they are context-bound and mining deep sentiment. The context of yellow for the words paper/sun/snow can vary greatly, in some cases it will be pleasant, in others pejorative.

When we put together the segments on which we are looking for hapaxes, it will no longer be a hapax. We only work with the extracted segments of 6250 words, because this has been empirically shown to be necessary for a completely reliable authorship attribution.

We identify the linguistic content profile of the author, i.e., preferred content, emotional, attitudinal stuff, so that we can do psychological, attitudinal profiling of them. We track the persistence of hapaxes, how they hold across time, which is why we need to compare old and new utterances, to track the evolution of the hapax list.

Hapax Analysis of Berlusconi's Speeches

The practical part of this thesis consists in gathering data and analyzing them. Our team is gathering data for a number of different languages (English, Czech, German, and others). I will focus on Italian. In order to have the best quality data and minimize bias as much as possible, we need data from different periods and utterances of the same speaker. These utterances need to be unprepared (preparation means priming for certain words that may be suggested by the dictionary and thus not make part of the speaker's active repertoire in comparison to other words).

Probably the most distinctive spoken utterances that are also easily accessible in Italian come from the politician Berlusconi. Transcripts of his speeches from publicly available recordings, which add up to many tens to hundreds of hours, are a possibility. I made transcriptions of 5 videos. To achieve a word count sufficient for three segments of 6250 words, seven of Berlusconi's speeches were used, of which five were my own transcriptions and two were taken from Bolasco's corpus (that was created for the book Parole in libertà: Un'analisi statistica e linguistica, 2006).

However, there is a corpus of his speeches already existing, which I have received from Sergio Bolasco. The corpus contains 111 speeches by Silvio Berlusconi, with a total length of 328,986 words (excluding headings), i.e. 325595 tokens, 19573 types.

In the corpus there are headings "parlato-scritto" and "scritto-parlato," indicating whether the speech was delivered on the basis of a prepared written text or whether it was at first orally presented and rewritten afterwards. Of course, it must still be taken into account that even speeches that were first delivered orally and for which there was no written draft may have been prepared or planned at least to some extent. In the case of higher-level political speeches, it is hard to imagine zero preparation (or at least forethought) of any speech, at least forethought about the target audience. Getting used to monotonous speeches in an ever-similar style reduces the amount of time needed to prepare the speech. In this case, the words and expressions used may become fixed. However, it would be very difficult to quantify this "level of preparedness"

[23] Faltýnek 2021

because we do not know the history of Berlusconi's preparations. There is also the danger that Berlusconi's speeches were prepared by someone else, one of his assistants or aides.

Thus, two analyses can be made: we have a large amount of data, but at the same time it is not possible to process such a large amount of data in detail. The large-scale analysis will settle for the quality of data that is available and bet on quantity over quality. The amount of data available is a huge advantage for us because we can track hapax variation across years and thus get a strong evidence for the "hapaxness" of a large number of words across many different utterances. A certain disadvantage is that these are all speeches of the same type, they are all political speeches, and we lack transcriptions of how Berlusconi speaks at home, with family and friends. Another disadvantage is the time-consuming nature of processing such data.

The manifestations are chronologically ordered in the corpus, which will allow diachronic observation of changes and developmental psychological profiling (done by other members of the team).

Due to the high number of hapaxes obtained, it might be possible to cluster hapaxes into groups and possibly observe the transformations of these groups in time. In particular, however, we will be interested in synchronic sentiment analysis to begin with. Interesting conclusions could also be drawn from the data from the percentage overview of the lexical classes of hapaxes, what names occur among them, from the comparison of the tagged list of hapaxes with the whole corpus, and from the analysis of the POS distribution.

The corpus also makes use of the three-dot marking, which seems to mean a pause, a hesitation. When we look at his manifestations, he is sometimes adding his typical smile. It is useful to make use of this marking and explore whether hesitation has any role around hapaxes.

There are also paragraphs created in the corpus, these are probably made according to some logical structuring of the speech.

After removing the headers from the corpus, we need to get segments of 6250 words, so we automatically split the corpus using the GSplit 3 file splitter. To achieve this, we first need to have the individual words on separate lines, which can be quickly achieved using regular expressions. In the splitter we then have the option to set the counter just per line. This gives us 52 full-length segments and a third of insufficient length.

In addition, an exploration of the recurrent lexis around hapaxes will be relevant, including a grammatical reflection on what the word attracts. If it attracts the same structures repeatedly, this will be a significant result. In addition to a qualitative, manual analysis (which includes subjective reflection and subjective sentiment analysis), it would be useful to do a frequency analysis of the surrounding area. We will care about how many times something appears at hapaxes compared to how many times it appears outside of hapaxes. In other words, whether it is a mean that highlights, emphasizes, or directly invokes the thing in question. For this purpose, WordCloud visualization (e.g., a tool from MonkeyLearn) can be used to compare the differences between the individual occurrences of a superhapax. On the other hand, insights from studying hapaxes can be used to improve WordCloud technology.

It is necessary to examine the overall thematic structure and compare it with the hapax structure. Does an individual talk about given topics all the time, or is a given hapax tied to a given topic? A thematic analysis, a list of themes, should be created from the entire text and compared to the

list of themes appearing in the hapaxes. Individual themes are interwoven across hapaxes, and individual hapaxes may be tied to more frequently occurring themes. The analysis could be done on even segments of length 5000 and compared, looking at the percentage of topic representation per word, for triples this is problematic, the more instances, the more certain the topic really attracts the word.

Results

Large-scale analysis

The Berlusconi corpus was segmented and hapaxes were obtained for each segment. The numbers of superhapax occurrences across segments were measured. The list of superhapaxes – both super hapax legomena and super dis legomena is attached in the appendix. It contains some interesting content and function words.

The sentiment analysis of a selected sample of these superhapaxes has not yet been completed (remains for further publications), only the hapax neighborhood extraction has been conducted through the software that is being developed by the hapax team. An interesting finding when doing sentiment analysis in small-scale analysis was the ideal amount of surrounding context needed to understand the meaning of the hapax. This was 150 words before and 20 words after the hapax. This would mean that the context would cause the use of the hapax (and not the other way around, that the hapax would trigger the context).

Small-scale analysis

From the three segments we get two dis legomena that occurred in all three segments and 34 hapax legomena that occurred in all three segments.

	3 SEGMENTS OF BERLUSCONI'S SPEECHES
SUPER DIS LEGOMENA	potere, realizzare
SUPER HAPAX LEGOMENA	assoluta, conosce, privata, serve, ritenuto, uffici, portare, impedire, chiare, vostra, progetto, vivere, ben, continuano, certa, guardando, presenza, pare, interesse, corso, economiche, vent, maggiore, miliardo, avremo, soldi, invito, cure, basta, pubbliche, stampa, sappiamo, qual, livelli

Table 1: Overview of super hapax legomena and super dis legomena in 3 analyzed segments of Berlusconi's speeches

For all of these 36 items, sentiment analysis was conducted. We will show a detailed example of what the analysis of one of the dis legomena, "potere" ("power") looked like.

POTERE. This word form can be either the infinitive form of the verb "can," which is often shortened to "poter" (poter has 4 occurrences in total in all segments) or the noun "power." In the first segment we have 4 more occurrences of the plural "poteri."

Here is the list of contextual meaning analysis of the individual occurrences:

1. disagreement with the left that man is subject to the state, we have a western tradition here; he who is in power serves

2. disagreement with the left, that man is subject to the state

3. reference to Marx, fear of the asymmetry of forces in Italy

4. as a verb: in order to cooperate with eastern countries – he talked about Chinese communism shortly before, these countries should get used to our values

5. purchasing power in relation to the lira, when the lira was abolished, half the purchasing power was lost

6. out of desperation Berlusconi ran for the office again because the communists were coming to power

As we can see, around this hapax, "potere" we have a constant recurrence of the theme of disagreement with the left. In one of his speeches, Berlusconi told a story from primary school, when a Russian priest came to lecture them on what it was like in Russia under Stalin. From then on, Berlusconi said, he was afraid of communism.

The rest of the sentiment analysis is included in the appendix.

However, here we still provide an overview of the themes that appeared in the sentiment analysis and the hapaxes associated with them, at the moment, this is a qualitative analysis that has been formalized only minimally.

DISAPPROVAL OF THE LEFT/COMMUNISM, FEAR OF COMMUNISM: POTERE 5/6, REALIZZARE 1/6, ASSOLUTA 1/3, PRIVATA 1/3, RITENUTO 1/3, PARE 1/3, CORSO 1/3, ECONOMICHE 1/3, MAGGIORE 1/3

OPPOSITION TO STATISM, WHICH SUPPRESSES FREEDOM: PRIVATA 1/3, PARE 1/3

PARTY VALUES: REALIZZARE min. 3/6 + close themes by other 2, CHIARE 2/3 where the third is somewhat close, PROGETTO 1/3, VIVERE 1/3, SAPPIAMO 1/3

TAXES, BENEFITS, POVERTY: ASSOLUTA 1/3, PRIVATA 1/3, SERVE 1/3 and the other one is close, PROGETTO 1/3, MILIARDO 1/3, CURE 1/3

PRISONS – FINANCING: ASSOLUTA 1/3

DUTY (ESPECIALLY IN CONNECTION WITH RUNNING IN ELECTIONS AGAINST THE COMMUNISTS, BUT ALSO AS A STATESMAN), SERVICE TO THE PEOPLE, TO THE STATE, HELP: CONOSCE 3/3, SERVE 1/3, RITENUTO 1/3, VOSTRA 1/3, BEN 1/3 (as a leader), CONTINUANO 1/3, GUARDANDO 1/3, INTERESSE 1/3, CORSO 1/3 (as a leader), ECONOMICHE 1/3, MAGGIORE 1/3, INVITO 2/3, CURE 1/3

GO VOTE AND CHOOSE A LEADER (ME, BERLUSCONI): INTERESSE 1/3, BASTA 1/3

PROGRAMME, BELIEF IN THE COUNTRY: SERVE 1/3, UFFICI 2/3, VOSTRA 2/3, CONTINUANO 1/3, INTERESSE 1/3

UNFAIRNESS OF THE STATE, TOO BUREAUCRATIC, TOO SPRAWLING: RITENUTO 1/3, PORTARE 1/3, CHIARE 1/3, PROGETTO 1/3, VIVERE 1/3, BEN 1/3, PRESENZA 1/3, ECONOMICHE 1/3, BASTA 1/3, PUBBLICHE 2/3

GOVERNMENT DEBT: UFFICI 1/3, IMPEDIRE 1/3

SENSITIVITY TO COMPLIANCE: PORTARE 1/3, CERTA 1/3, SAPPIAMO 1/3, QUAL 1/3

FUTURE: REALIZZARE 1/6, IMPEDIRE 1/3, VENT 1/3

MAFIA: PROGETTO 1/3, GUARDANDO 1/3, PARE 1/3, CORSO 1/3

ACCUSED/JUSTIFICATION: VIVERE 1/3, BEN 1/3, PARE 1/3, BASTA 2/3, STAMPA 2/3

LIRA: CONTINUANO 1/3

CURRENCES: MILIARDO 1/3

XENOPHOBIA: CERTA 1/3

FINDING A COMMON AGENDA: CERTA 1/3, AVREMO 1/3, INVITO 1/3, QUAL 2/3

FOOLISHNESS = VISIONARISM: GUARDANDO 1/3, SOLDI 1/3, PUBBLICHE 1/3

DEFAMATION OF ITALY: PRESENZA 1/3, MAGGIORE 1/3

CRIME, SECURITY: PRESENZA 1/3, MAGGIORE 1/3

IMMIGRATION: VENT 2/3, MILIARDO 1/3, AVREMO 1/3, SODLI 1/3, LIVELLI 1/3

UNEMPLOYMENT: MAGGIORE 1/3, CURE 1/3, SAPPIAMO 2/3, LIVELLI 1/3

FATHER'S FINANCIAL RESPONSIBILITY: AVREMO 1/3, SOLDI 1/3

As can be seen, some topics could perhaps be merged, while others are too broad. E.g. DUTY, HELP look too broad, but the two topics are so intertwined that it is not easy to separate them.

If we take the hapax "POTERE" again and look at the thematic and sentiment analysis, what might have looked as promising, looks precarious now. The fact that "poter" has 4 occurrences in total in all segments and that in the first segment we have 4 more occurrences of the plural "poteri," may be raising the question of lemmatization. What looks somewhat promising, is that 5 cases of "POTERE" are related to the topic of disagreement with the left, while other two super hapaxes appear within the context of this topic too, but only once.

Now it would be useful to quantitatively analyze the descriptions obtained, and to do a topic and hapax structure of the text, in order to know whether the author talks about these topics all the time or whether a given hapax is tied to a given topic. It would be a good idea to measure how many times something appears around superhapaxes versus somewhere off – is it something that either just highlights or emphasizes or directly invokes the phenomenon?

Conclusion

The paper attempted to provide a categorization of hapaxes and focused on evaluating the impact of the discovery of the so-called superhapax, a low-frequency item with a regular occurrence, on ideas about the mental lexicon. Furthermore, the possible roles of hapaxes and superhapaxes in language were briefly described: for instance, in information theory, synergetic and quantitative linguistics, their economizing and sentimental function, or their role of cohesive means in grammar and syntax.

From the study of literature, the necessity of studying memory was found to be crucial, and the relevant phenomena associated with memory functioning were identified, which quality mental lexicon models must reflect. These include in particular the frequency effect and the contiguity effect, which is associated with contextual priming, contextual learning, and a phenomenon observed in hapax research called the Bruntál effect. Very simple ideas have been described about the workings of the brain and the different types of memory, which are in general thought to be separate – but it seems that they may not be completely separate. The description also touched on priming.

The most significant finding was the concept of syntagmatic-paradigmatic and syntagmatic-paradigmatic-syntagmatic shift as described by Petrey (1977). Children show a preference for syntagmatic associations, e.g., add-flour, whereas adults show a preference for paradigmatic associations, e.g., table-chair. However, for some expressions, their usage becomes so solidified over time that the speaker reverts to a syntagmatic association, e.g. hermetically-sealed. This could explain the behavior and functioning of the superhapaxes that are associated with the given topic.

Next, the different models of the mental lexicon were discussed in the light of the hapax evidence.

The hierarchical network model and the spreading activation model (which, while acknowledging the individual organization of items, otherwise more or less fails to deal with important aspects of language processing and may be based on inappropriate phenomena) have been criticized.

The assumptions of the adaptive nature of mind model were also slightly criticized because of poor hapax individuation.

The possibility (advantages and disadvantages) of using WordNets (and similar models) to model the mental lexicon was discussed.

A possible explanation of the (n)ROUSE model's operation via hapax registration was given.

For models of lexical access, the role of focus and bias in making certain aspects of language parsable was considered. The cohort model was criticized because it proposes something that cannot be expected on the basis of hapax evidence, since we observe rather semantic relations and a conceptually unguided, non-normative vocabulary, in contrast to a cohort of phonologically identical word-initial words. The use of the TRACE model was considered. The need to include the phenomenon of S-P-S shift in the connectionist representations was discussed.

Multilingual models could not be omitted; the BIA and BIA+ models were discussed in detail. Additional hapax data are needed to evaluate these models and are planned. Word learning is expected to be context-dependent: by priming the context in which a word is frequently encountered, we can recall that word. However, it is possible that the hypotheses used by these models are flawed.

An important, very attractive model that was explored was the multiplex model. This involves explosive learning, where the network is connected through the emergence of important nodes representing abstract words. It is structured into four layers, the inclusion of free associations is particularly satisfying (we can explain their causes). In the context of this model, we encounter the need to investigate the development of hapaxes from infancy onwards and to observe what words become superhapaxes. In any case, this is a workable model to which we could add work with emotions.

Another point of the paper is a preliminary proposal of the possibility of a prototype model of a mental lexicon that would explain the concept of superhapax and its Bruntál effect and S-P-S shift. The hypothesis of a cognitive limit on the number of actively maintained connections with words seems particularly interesting to test. The nature of the empirically obtained number 6250 as the segment length needed to confidently identify the author has been discussed. Along

with a verification of contextual word learning (and thus a refutation of the independence of semantic memory), experimental designs were offered.

Finally, hapax analyses of the Berlusconi corpus and sentiment analysis (with significant results) were performed on three segments of 6250 words. A hapax analysis of the whole segmented corpus of 320 thousand words was also performed. Some words appear to be topically bound, but for the majority no such trend was found. This may be due to the nature of the material, where it is purely political speeches. Thus, larger and more detailed sentiment analyses are needed to see whether what seems to work on other languages (such as English) also works on Italian.

Literature and resources

Altmann, Gerry T.M. (1997). "Words, and how we (eventually) find them." The Ascent of Babel: An Exploration of Language, Mind, and Understanding. Oxford: Oxford University Press. pp. 65–83

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist, 51*(4), 355-365.

Atkinson, R. C. and Shiffrin, R. (1968). M. The Psychology of Learning and Motivation, 2, 89-195. 1968.

Baddeley AD (November 1966). "Short-term memory for word sequences as a function of acoustic, semantic and formal similarity" (PDF). Quarterly Journal of Experimental Psychology. 18 (4): 362–5. doi:10.1080/14640746608400055. PMID 5956080. S2CID 32498516.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol. 8, pp. 47–89). New York: Academic Press.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. Journal of Verbal Learning and Verbal Behavior, 14(6), 575–589. doi:10.1016/s0022-5371(75)80045-4

Barbaresi, A. (2014). Language-classified Open Subtitles (LACLOS): download, extraction, and quality assessment. Ph.D. thesis, Last Accessed: 15 January 2017. BBAW, URL https://hal.archives-ouvertes.fr/hal-01083746/document

Bednaříková, Božena. (2011). Towards (Proto)typing of Morphological Processes. In: Czech and Slovak Linguistic Review 1/2011. Olomouc: Palacký University

Blackmer, Elizabeth R. and Mitton, Janet L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. Cognition 39: 173–194.

Bock, Kathryn and Willem Levelt. (1994). Language Production: Grammatical Encoding. In: Gernsbacher (ed.): Handbook of Psycholinguistics. p. 945-984.

Bock M. (1986). The influence of emotional meaning on the recall of words processed for form or self-reference. Psychol Res. 48:107—112

Bolasco, Sergio, Luca Giuliano & Nora Galli de' Paratesi. (2006). Parole in libertà: Un'analisi statistica e linguistica. Roma: Manifestolibri.

Bosch-Bouju C, Hyland BI, Parr-Brownlie LC. (2013). Motor thalamus integration of cortical, cerebellar and basal ganglia information: implications for normal and parkinsonian conditions. Front Comput Neurosci. 2013 Nov 11;7:163. doi: 10.3389/fncom.2013.00163. PMID: 24273509; PMCID: PMC3822295.

Brendel, B., Erb, M., Riecker, A., Grodd, W., Ackermann, H., and Ziegler, W. (2011). Do we have a "mental syllabary" in the brain? An fMRI study. Mot. Control 15, 34–51. doi: 10.1123/mcj.15.1.34

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. Psychological Bulletin, 109(2), 204–223. doi:10.1037/0033-2909.109.2.204

Brysbaert, M., Warriner, A. B. & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. Behavior research methods 46, 904–911 (2014).

Brysbaert M, Stevens M, Mandera P and Keuleers E (2016). How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. Front. Psychol. 7:1116. doi: 10.3389/fpsyg.2016.01116

Butterworth, B. (1989). "Lexical access in speech production," in Lexical Representation and Process, ed. W. Marslen-Wilson, (Cambridge, MA: MIT Press), 108–135.

Caramazza, A. (1997). How many levels of processing are there in lexical access? Cogn. Neuropsychol. 14, 177–208. doi: 10.1080/026432997381664

Cholin, J. (2008). The mental syllabary in speech production: an integration of different approaches and domains. Aphasiology 22, 1127–1141. doi: 10.1080/02687030701820352

Collins, A. M. and Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior*, 8(2), 240-247. doi:10.1016/S0022-5371(69)80069-1

Collins, A. M. & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. Psychological review 82, 407.

Coltheart, M. (1981) The MRC psycholinguistic database. The Quarterly Journal of Experimental Psychology 33, 497–505.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences, 24(1), 87-114. doi:10.1017/S0140525X01003922

Craik, F.I.M, & Lockhart, R.S. (1972). Levels of processing: A framework for memory research. Journal of Verbal Learning and Verbal Behavior, 11, 671-684.

De Deyne, S., Kenett, Y. N., Anaki, D., Faust, M. & Navarro, D. J. (2016). Large-scale network representations of semantics in the mental lexicon. In Big data in cognitive science: From methods to insights 174–202 Psychology Press: Taylor & Francis.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. Psychol. Rev. 93:283. doi: 10.1037/0033-295x.93.3.283

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., and Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. Psychol. Rev. 104:801. doi: 10.1037/0033-295x.104.4.801

Dijkstra, T. & Van Heuven. (1998). The BIA model and bilingual word recognition. In J. Grainger & A.M. Jacobs. (Eds.), Localist connectionist approaches to human cognition (189-225). Mahwah, NJ: Erlbaum.

Dijkstra, T., Van Heuven, W.J.B., & Grainger, J. (1998). Simulating cross-language competition with the bilingual interactive activation model. Psychologica Belgica, 38, 177-196.

Dijkstra, A. F. J., & Heuven, W. V. (2002). The architecture of the bilingual word recognition system: From identification to decision. Bilingualism: Language and Cognition, 5, 175-197.

Duncan, Carl. (2013). Retrieval of low-frequency words from mixed lists. Bulletin of the Psychonomic Society. 4. 137-138. 10.3758/BF03334222.

Eichenbaum H, Cohen NJ (1993). Memory, Amnesia, and the Hippocampal System. MIT Press.

Entwistle, D. R. (1966). Word associations of young children. Baltimore: The Johns Hopkins Press.

Faltýnek, Dan a Vladimír Matlach. (2021). Hapax remains: Regularity of low-frequency words in authorial texts. In: Digital Scholarship in the Humanities.

Faltýnek, Dan, Ludmila Lacková a Hana Owsianková. (2020). Once again about the hapax grammar: Epigenetic linguistics. In: Linguistic Frontiers. DOI: 10.2478/lf-2019-0002

Faltýnek, Dan a Vladimír Matlach. (2020). Hapax Remains: authorial features of textual cohesion in authorship attribution (Preprint).

Faltýnek, Dan. (2020). It will certainly be found that some words are literally repeated: Horecký's hypersyntax. In: Jazykovedny Casopis 71 (2), 185-196.

Faltýnek, Dan. (2021). Text Structure Chance and Necessity: the Regularity of the Occurrence of Low-frequency Words in Authorial Texts.

Faltýnková, Klára. (2022). Visualization of the Bruntál effect (painting). Olomouc.

Fengxiang F. (2010): An Asymptotic Model for the English Hapax/Vocabulary Ratio. Computational Linguistics, 36, 4, p. 631–637.

Finn, P. J. (1977-1978). Word frequency, information theory, and cloze performance: A transfer feature theory of processing in reading. Reading Research Quarterly, 13(4), 508–537. https://doi.org/10.2307/747510

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E.Walker (Eds.), New approaches to language mechanisms (pp. 257–287). Amsterdam: North-Holland.

Freud, S. (1901). Psychopathology of Everyday Life (Translated by Bell, A.). London: Penguin Books Ltd.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception, Journal of Experimental Psychology: Human Perception and Performance. 6, I IO- 125

Garrett, M. (1980). “Levels of processing in sentence production,” in Language Production, ed. B. Butterworth, (London: Academic Press).

Glanzer M, Adams JK. (1985). The mirror effect in recognition memory. Memory & Cognition. 1985;12:8–20.

Gluck M, Mercado E, Myers C (2014). Learning and Memory From Brain to Behavior Second Edition. New York: Kevin Feyen. p. 416. ISBN 978-1429240147

Gollan, Tamar & Slattery, Timothy & Goldenberg, Diane & Van Assche, Eva & Duyck, Wouter & Rayner, Keith. (2011). Frequency Drives Lexical Access in Reading but Not in Speaking: The Frequency-Lag Hypothesis. Journal of experimental psychology. General. 140. 186-209. 10.1037/a0022256.

Griffin Z. M. (1999). Frequency of meaning use for ambiguous and unambiguous words. Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc, 31(3), 520–530. https://doi.org/10.3758/bf03200731

de Groot AMB. (1989). Representational aspects of word imageability and word frequency as assessed through word association. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1989;15:824–845.

Hallig, R. and W. von Wartburg (1952). Begriffssytem als Grundlage für die Lexikographie. Versuch eines Ordnungsschemas. Berlin: Akademie-Verlag.

Hansen, S. J., McMahon, K. L., & de Zubicaray, G. I. (2019). The neurobiology of taboo language processing: fMRI evidence during spoken word production. Social cognitive and affective neuroscience, 14(3), 271–279. https://doi.org/10.1093/scan/nsz009

van Heuven, W.J.B., Dijkstra, T., & Grainger, J. (1998). "Orthographic Neighborhood Effects in Bilingual Word Recognition." Journal of Memory and Language. pp. 458-483.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A. & Smith, L. (2009). Longitudinal analysis of early semantic networks preferential attachment or preferential acquisition? Psychological Science 20, 729–739

Holoyak, K. J., & Glass, A. L. (1975). The role of contradictions and counterexamples in the rejection of false sentences. Journal of Verbal Learning & Verbal Behavior, 14(2), 215–239. https://doi.org/10.1016/S0022-5371(75)80066-1

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. In: Journal of Mathematical Psychology, 46, 269e299.

HUBER, D. E., SHIFFRIN, R. M., LYLE, K. B., &RUYS, K. I. (2001). Perception and preference in short-term word priming. Psychological Review,108, 149–182.

D.E. Huber, R.C. O’Reilly. (2003). Persistence and accommodation in short-term priming and other perceptual paradigms: Temporal segregation through synaptic depression. Cognitive Science, 27 (2003), pp. 403-430

Hulme, Charles & Stuart, George & Brown, Gordon & Morin, Caroline. (2003). High- and low-frequency words are recalled equally well in alternating lists: Evidence for associative effects

in serial recall. Journal of Memory and Language. 49. 500-518. 10.1016/S0749-596X(03)00096-2.

Indefrey, P & Levelt, W. (2004). Indefrey, P. & Levelt, W.J.M. The spatial and temporal signatures of word production components. Cognition 92, 101-144. Cognition. 92. 101-44. 10.1016/j.cognition.2002.06.001.

Indefrey, P. (2011). The spatial and temporal signatures of word production components: a critical update. Front. Psychol. 2:255. doi: 10.3389/fninf.2013.000255

Kerouac, Jack. (1957). On the Road. New York: Viking Press.

Keuleers, E., Lacey, P., Rastle, K. & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic english words. Behavior Research Methods 44, 287–304 (2012).

Krashen, Stephen (1977). "Some issues relating to the monitor model." In Brown, H; Yorio, Carlos; Crymes, Ruth (eds.). Teaching and learning English as a Second Language: Trends in Research and Practice: On TESOL '77: Selected Papers from the Eleventh Annual Convention of Teachers of English to Speakers of Other Languages, Miami, Florida, April 26 – May 1, 1977. Washington, DC: Teachers of English to Speakers of Other Languages. pp. 144–158. OCLC 4037133.

Krashen, S. (1982). Principles and practice in second language acquisition. Oxford: Pergamon Press "Archived copy." Archived from the original on July 16, 2011. Retrieved November 25, 2010.

Krashen, S.D. (1989). We acquire vocabulary and spelling by reading: additional evidence for the input hypothesis, Modem Language Journal, vol. 73, nº4, pp. 440–464

Kröger, B. J., and Cao, M. (2015). The emergence of phonetic–phonological features in a biologically inspired model of speech processing. J. Phon. 53, 88–100. doi: 10.1016/j.wocn.2015.09.006

Kuperman, V., Stadthagen-Gonzalez, H. & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. Behavior Research Methods 44, 978–990.

LaBar, K. S.; Phelps, E. A. (1998). "Arousal-mediated memory consolidation: Role of the medial temporal lobe in humans." Psychological Science. 9 (6): 490–493. doi:10.1111/1467-9280.00090. S2CID 15003037.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. Cognitive Psychology, 6(2), 293–323. doi:10.1016/0010-0285(74)90015-2

Lakoff, G. & Johnson, M. (1980). Metaphors we live by. Chicago: University of Chicago Press.

Lakoff G. (1990). Women, Fire, and Dangerous Things: What Categories Reveal About the Mind. University Of Chicago Press.

Lashley, Karl S. (1951). The problem of serial order in behavior. In L. A. Jeffress, ed., Cerebral Mechanisms in Behavior, pp. 112–146. New York: Wiley.

Laufer, Batia. (1998). The Development of Passive and Active Vocabulary in a Second Language: Same or Different?. Applied Linguistics. 19. 10.1093/applin/19.2.255.

Legendre, Géraldine, Miyata, Yoshiro & Smolensky, Paul. (1990). Harmonic Grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. In Proceedings of the twelfth annual conference of the Cognitive Science Society (pp. 388–395). Cambridge, MA: Lawrence Erlbaum. Report CU-CS-465-90. Computer Science Department, University of Colorado at Boulder.

Lemke, Harrison [@hplemke]. (2018, December 14). What do you call a lego piece that appears only once in a given set? – A hapax legomenon. [Tweet]. Twitter. https://twitter.com/hplemke/status/1073609922569560066

Lerner, G. (2013). On the place of hesitating in delicate formulations: A turn-constructional infrastructure for collaborative indiscretion. In M. Hayashi, G. Raymond, & J. Sidnell (Eds.), Conversational Repair and Human Understanding (Studies in Interactional Sociolinguistics, pp. 95-134). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511757464.004

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. Cognition, 14, 41–104. DOI: https://doi.org/10.1016/0010-0277(83)90026-4

Levelt, W. J. M. (1989). Speaking: From intention to articulation. MIT Press.

Levelt, W. J. M., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. Behav. Brain Sci. 22, 1–75.

Lucas M. (1999). Context effects in lexical access: a meta-analysis. Mem Cognit. 1999 May; 27(3): 385-98. doi: 10.3758/bf03211535. PMID: 10355230.

Lucas, M. (2000). Semantic priming without association: a meta-analytic review. Psychon. Bull. Rev. 7, 618–630. doi: 10.3758/bf03212999

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. Ear and hearing, 19(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Marslen-Wilson, W.D., & Welsh, A.B. (1978). Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology, 10, 29-63.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition*, 25, 71-102.

Matoré, G. (1953). La méthode en lexicologie. Paris: Didier.

McClelland, L. L., & J. Elman. (1985). The TRACE model of speech perception. Cognitive Psychology: Volume 18, Issue 1, January 1986, Pages 1-86

McGregor, K. K., and Waxman, S. R. (1998). Object naming at multiple hierarchical levels: a comparison of preschoolers with and without word-finding deficits. J. Child Lang. 25, 419–430. doi: 10.1017/s030500099800347x

Miller, G. A. (1956). "The magical number seven, plus or minus two: Some limits on our capacity for processing information." Psychological Review. 63 (2): 81–97. CiteSeerX 10.1.1.308.8071. doi:10.1037/h0043158

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM 38, 39-41

Morris, R.K. (1992). Sentence Context Effects on Lexical Access. In: Rayner, K. (eds) Eye Movements and Visual Cognition. Springer Series in Neuropsychology. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-2852-3_19

Ogden, C. K. (1954). THE BASIC WORDS, A DETAILED ACCOUNT OF THEIR USES. Kegan Paul & Co Ltd.

Olson, IR; Plotzker, A; Ezzyat, Y (2007). "The enigmatic temporal poles: A review of findings on social and emotional processing." Brain. 130 (7): 1718–1731. doi:10.1093/brain/awm052. PMID 17392317.

Packard, Jerome L (2000). "Chinese words and the lexicon." The Morphology of Chinese: A Linguistic and Cognitive Approach. Cambridge: Cambridge University Press. pp. 284–309.

Pajek Datasets. (2003). The Edinburgh Associative Thesaurus [online]. Pajek Data: The Edinburgh Associative Thesaurus. [accessed May 4th 2022] http://vlado.fmf.uni-lj.si/pub/networks/data/dic/eat/Eat.htm#:~:text=The%20Edinburgh%20Associative%20Thesaurus%20(EAT,)%2C%20but%20empirical%20association%20data.

Palermo, D. S., & Jenkins, J. J. (1964). Word association norms: Grade school through college. U. Minnesota Press.

Park, H., Reder, L. M., & Dickison, D. (2005). The effects of word frequency and similarity on recognition judgments: the role of recollection. Journal of experimental psychology. Learning, memory, and cognition, 31(3), 568–578. https://doi.org/10.1037/0278-7393.31.3.568

Popescu, Ioan-Iovitz and Gabriel Altmann. (2008). Hapax Legomena and Language Typology. In: Journal of Quantitative Linguistics 2008, Vol 15, Number 4, pp. 370-378. doi:10.1080/09296170802326699

Raaijmakers, Jeroen G. W.; Shiffrin, Richard M. (1981). "Search of associative memory." Psychological Review. 88 (2): 93–134. doi:10.1037/0033-295X.88.2.93.

Rosch, E.H. (1973). "Natural categories." Cognitive Psychology. 4 (3): 328–50. doi:10.1016/0010-0285(73)90017-0.

Rosch, E.H. (1975). "Cognitive reference points." Cognitive Psychology. 7 (4): 532–47. doi:10.1016/0010-0285(75)90021-3. S2CID 54342276.

Rosch, E.H. (1975). "Cognitive representation of semantic categories," Journal of Experimental Psychology 104(3): 192-233.

Rosch, E.H.; Mervis, C.B.; Gray, W.D.; Johnson, D.M.; Boyes-Braem, P. (1976). "Basic objects in natural categories." Cognitive Psychology. 8 (3): 382–439. CiteSeerX 10.1.1.149.3392. doi:10.1016/0010-0285(76)90013-X. S2CID 5612467.

Mervis, C.B.; Rosch, E. (1981). "Categorization of Natural Objects." Annual Review of Psychology. 32: 89–113. doi:10.1146/annurev.ps.32.020181.000513.

Radanovic M, Azambuja M, Mansur LL, Porto CS, Scaff M. (2003). Thalamus and language: interface with attention, memory and executive functions. Arq Neuropsiquiatr. 2003 Mar;61(1):34-42. doi: 10.1590/s0004-282x2003000100006. Epub 2003 Apr 16. PMID: 12715016.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. Journal of Verbal Learning & Verbal Behavior, 12(1), 1–20. https://doi.org/10.1016/S0022-5371(73)80056-8

Salvato G, Scarpa P, Francione S, Mai R, Tassi L, Scarano E, Lo Russo G, Bottini G. (2016). Declarative long-term memory and the mesial temporal lobe: Insights from a 5-year postsurgery follow-up study on refractory temporal lobe epilepsy. Epilepsy Behav. 2016 Nov;64(Pt A):102-109. doi: 10.1016/j.yebeh.2016.08.029. Epub 2016 Oct 11. PMID: 27736656.

Sass, Katharina, Sören Krach, Olga Sachs & Tilo Kircher. (2009). Lion – tiger – stripes: Neural correlates of indirect semantic priming across processing modalities. In: NeuroImage Volume 45, Issue 1, March 2009, Pages 224-236.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. Psychological Review, 115, 893e912.

Seidenberg, Mark S, Michael K Tanenhaus, James M Leiman, Marie Bienkowski. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. In: Cognitive Psychology, Volume 14, Issue 4, 1982, Pages 489-537, ISSN 0010-0285, https://doi.org/10.1016/0010-0285(82)90017-2.

Schegloff, E. (2013). Ten operations in self-initiated, same-turn repair. In M. Hayashi, G. Raymond, & J. Sidnell (Eds.), Conversational Repair and Human Understanding (Studies in Interactional Sociolinguistics, pp. 41-70). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511757464.002

Shiffrin, Richard & Huber, David. (2001). ROUSE and the multinomial model: A priori versus a posteriori predictions.

Simpson, G. B. (1981). Meaning dominance and semantic context in the processing of lexical ambiguity. Journal of Verbal Learning and Verbal Behavior, 20, 120-136.

Slowiaczek, L. M., Nusbaum, H. C., & Pisoni, D. B. (1987). Phonological priming in auditory word recognition. Journal of experimental psychology. Learning, memory, and cognition, 13(1), 64–75. https://doi.org/10.1037//0278-7393.13.1.64

Smolensky, Paul. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. Artificial intelligence 46.1-2: 159-216.[1]

Steinbeck, John. (2013). O myších a lidech. Frýdek-Místek: Alpress.

Stella, M. & Brede, M. (2015). Patterns in the English language: Phonological networks, percolation and assembly models. Journal of Statistical Mechanics: Theory and Experiment 2015, P05006

Stella, M., Beckage, N. M. & Brede, M. (2017). Multiplex lexical networks reveal patterns in early word acquisition in children. Scientific Reports 7

Stella, Massimo, Nicole M. Beckage, Markus Brede and Manlio De Domenico. (2018). Multiplex model of mental lexicon reveals explosive learning in humans. Scientific Reports (2018) 8:2259, doi:10.1038/s41598-018-20730-5

Stemberger, J. P. (1985). An interactive activation model of language production. Prog. Psychol. Lang. 1, 143–186.

Stille, C. M., Bekolay, T., Blouw, P., & Kröger, B. J. (2020). Modeling the Mental Lexicon as Part of Long-Term and Working Memory and Simulating Lexical Access in a Naming Task Including Semantic and Phonological Cues. Frontiers in Psychology, 11. doi:10.3389/fpsyg.2020.01594

The k2p blog. (2017). There is a cognitive limit (the Wordsmith number) to the number of words you can know? [online] The k2p blog. [accessed April 26th 2022] https://ktwop.com/2017/08/05/there-is-a-cognitive-limit-the-wordsmith-number-to-the-number-of-words-you-can-know/#:~:text=just%20opinions-,There%20is%20a%20cognitive%20limit%20(the%20Wordsmith%20number)%20to%20the,a%20vocabulary%20of%2060%2C000%20words.

Thompson-Schill, S.L., D'Esposito, M., Aguirre, G.K., Farah, M.J., (1997). Role of the left inferior parietal cortex in retrieval of semantic knowledge: a reevaluation. Proceedings of the National Academy of Sciences of the United States of America 94, 14792-14797

Thorndike, E. L., & Lorge, I. (1944). The teacher's word book of 30,000 words. Bureau of Publications, Teachers Co.

Trier, Jost. (1973). Das sprachliche Feld. Eine Auseinandersetzung. In: Jost Trier: Aufsätze und Vorträge zur Wortfeldtheorie. hrsg. v. Anthony van der Lee und Oskar Reichmann, The Hague, Paris, S. 150–151.

Tronson, N. C.; Taylor, J. R. (2007). "Molecular mechanisms of memory reconsolidation." Nature Reviews Neuroscience. 8 (4): 262–275. doi:10.1038/nrn2090

Tulving, E. (1972). "Episodic and semantic memory." In Tulving, E.; Donaldson, W. (eds.). Organization of Memory. New York: Academic Press. pp. 381–402.

Unger Nina, Heim Stefan, Hilger Dominique I., Bludau Sebastian, Pieperhoff Peter, Cichon Sven, Amunts Katrin, Mühleisen Thomas W. (2021). Identification of Phonology-Related Genes and Functional Characterization of Broca's and Wernicke's Regions in Language and Learning Disorders. In Frontiers in Neuroscience 15, 2021 DOI=10.3389/fnins.2021.680762

Vakoch, D. A., & Wurm, L. H. (1997). Emotional connotation in speech perception: Semantic associations in the general lexicon. Cognition and Emotion, 11(4), 337–349. https://doi.org/10.1080/026999397379827

Van den Bussche, Eva & Van den Noortgate, Wim & Reynvoet, Bert. (2009). Mechanisms of Masked Priming: A Meta-Analysis. Psychological bulletin. 135. 452-77. 10.1037/a0015329.

Venezky, R. L., & Calfee, R. C. (1970). The reading competence model. In H. Singer & R. B. Ruddell (Eds.), Theoretical models and processes of reading. Newark, Delaware: International Reading Association.

Vitevitch, M. S., Chan, K. Y., and Roodenrys, S. (2012). Complex network structure influences processing in long-term and short-term memory. J. Mem. Lang. 67, 30–44. doi: 10.1016/j.jml.2012.02.008

van Wijk, Carel, and Kempen, Gerard. (1987). A dual system for producing selfrepairs in spontaneous speech: evidence from experimentally elicited corrections. Cognitive Psychology 19: 403–440.

WolframResearch. WordData source information. http://reference.wolfram.com/language/note/WordDataSourceInformation.html (last accessed: 2017-05-14).

Zhang, Jiawei. (2019). Basic Neural Units of the Brain: Neurons, Synapses and Action Potential. arXiv preprint arXiv:1906.01703

List of figures

List of tables

List of appendices

Bolasco, Sergio, Luca Giuliano & Nora Galli de' Paratesi. (2006). Corpus Berlusconi.

List of notions related to mental lexicons

A zip file containing: 3 segments used for small-scale analysis, list of hapax legomena from the 3 segments of small-scale analysis, list of dis legomena from the 3 segments of small-scale analysis, sentiment analysis of the 3 segments, list of hapax legomena for the whole corpus, list of dis legomena for the whole corpus