Studi e Ricerche

Studi umanistici – Ricerche sull'Oriente

Percorsi in Civiltà dell'Asia e dell'Africa II

Quaderni di studi dottorali alla Sapienza

a cura di
Marina Miranda

University Press

SAPIENZA
UNIVERSITÀ EDITRICE

Collana Studi e Ricerche 130

Studi umanistici
Serie Ricerche sull'Oriente

Percorsi in Civiltà dell'Asia e dell'Africa II

Quaderni di studi dottorali alla Sapienza

a cura di
Marina Miranda

SAPIENZA
UNIVERSITÀ EDITRICE
2023

Sapienza Università Editrice
Piazzale Aldo Moro 5 – 00185 Roma

www.editricesapienza.it
editrice.sapienza@uniroma1.it

Iscrizione Registro Operatori Comunicazione n. 11420
Registry of Communication Workers registration n. 11420

ISBN: 978-88-9377-260-0

DOI: 10.13133/9788893772600

Pubblicato nel mese di gennaio 2023 | *Published in January 2023*

Impaginazione a cura di | *Layout by*: Tonio Savina

In copertina | *Cover image*: foto di cmcderm1 da Istockphoto.com, ID 91629206.

Indice

5. Synonymy in Korean Lexicon through the Lens of Vector Semantics

Valeria Ruscio

5.1. Introduction: Korean lexicon

Sino-Korean vocabulary refers to Korean words of Chinese origin. Therefore, this vocabulary includes words that entered the Korean lexicon as loan-words from Chinese, as well as new Korean words created from Chinese characters (Song 2005). On the other hand, native-Korean words are words not loaned or calqued from Chinese or other languages.

Usually, native-Korean words designate fundamental ideas to basic human life and are usually associated with traditional Korean culture; in particular, they predominate within the particle class (words with grammatical functions) and among onomatopoeias or ideophones[1]. This is not unexpected because words that express human experience or grammatical functions are less likely to be borrowed from other languages. Sino-Korean words are typically used in formal or literary contexts and to express abstract or complex ideas (Choo, Kwak 2008).

The relation linking native-Korean to the Sino-Korean words is of synonymic nature; in fact, those two categories of words may represent two different ways of expressing the same concept. The extensive Chinese influence on Korean language resulted in a sizable number of Sino-Korean words (Byon 2017); in fact, Sino-Korean words constitute

[1] Ideophones are words depicting a sensory perception with their sound; so, they do not represent the sound itself of the depicted concept, but rather they attempt to render phonetically one of its properties (Akita, Pardeshi 2019).

about 60% of the South Korean vocabulary, native-Korean words and loanwords from other languages (mostly from English) compose the remaining 40%. The vastity of the loans from Chinese also influenced the Korean word formation mechanism[2], so that Sino-Korean vocabulary has continued to grow in South Korea, where Chinese characters assume other meanings and are used to produce new words in Korean that do not exist in Chinese (Song 2005).

5.2. Methodology

The purpose of this study is to analyze the concept of synonymy and polysemy in the Korean lexicon, specifically by studying the relationship between native-Korean and Sino-Korean lexicon. This will be carried out by using computational linguistics techniques mentioned below, and specifically, due to the field of study called vector semantics, it is possible to compare words numerically and obtain measurable similarity results.

5.3. Synonymy

A synonym for a word *w1*, is a word *w2* that has the same or nearly the same meaning as *w1* in the same language (Oxford References)[3]. We can say that two words are synonyms if they have the same propositional meaning, that is to say, if a word can replace another in any sentence without changing the truth conditions[4] of the sentence (Jurafsky, Martin 2014). There are other types of semantic relations between words, such as word similarity, or the words can also belong to the same semantic field – or frame. Synonymy is not the only type of

---

2 The Word Formation Mechanism is the process via which new words are produced in a language; this can be done by modifying existing words, or by creating completely new words (Shi 2015).

3 See <https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100547710>.

4 A truth condition is the condition under which a sentence is true; and if the truth conditions for the sentence are clear, the meaning of a sentence is correctly conveyed (Birner 2013).

relationship existing between words; it is, in fact, one of the rarest. This because, according to the principle of contrast «any difference in form in a language marks a difference in meaning» (Jurafsky, Martin 2014), and generally a pair of words that happen to have a similar meaning may also have some differences. It is also one of the reasons why it is useful to consider similarity measures. For example, if we consider a pair of words such as "window" and "door", despite not being synonyms at all, they still belong to the same semantic field. Usually, they are associated and perceived as similar because i) both describe an object that can be opened and closed, representing a "passage" from the inside of a building to the outside, ii) both have a handle. As a matter of fact, it is possible to find these two words in very similar contexts.

5.4. Vector semantics

«The idea of vector semantics is thus to represent a word as a point in some multidimensional semantic space» (Jurafsky, Martin 2014). Vector semantic is a field of linguistics that gained popularity towards the end of the 20th century, starting from Wittgenstein's assertion that «the meaning of a word is its use in the language», as opposed to the field of linguistics that theorized context-free grammar, saying that words cannot be defined by logical, deterministic rules, but must be defined by how people use them (Jurafsky, Martin 2014).

«You shall know a word by the company it keeps» (Firth, 1957). Thus, linguists of the distributionalist school had an insight: to define a word by its context or by its distribution in a specific language. The distribution is the set of contexts where the word occurs, the words they co-occur with, and the grammatical environment where it can be found. So, two words occurring in the same distributions are likely to have similar meanings (Jurafsky, Martin 2014).

We can, therefore, explain vector semantics through two main ideas: the first is related to the distributionalist intuition, namely, that words can be defined by looking at the words that co-occur with them, and the second concerns the fact that a word $\underline{w}$ can be defined as a vector (i.e., as a list of numbers representing the coordinates of a point in an N-dimensional space).

Usually, vectors used to represent words are called word

embedding, since the word is "embedded" or projected into a vector space (Jurafsky, Martin 2014).

5.5. Vector Space Model and embeddings

The Vector Space Model is an algebraic model for representing objects as vectors. Those objects can be words, documents, sentences, concepts, entities, or any element with its own semantics and for which it makes sense to define a notion of similarity. In the vector space model, objects are represented in a continuous multidimensional space. In the context of computational linguistics, the space we refer to is the semantic space, and representations of objects in this space are called distributed representations (Pilehvar, Camacho-Collados 2020).

To be able to calculate the similarity between two words, it is necessary to consider their vector representations, and to obtain them, two different types of architectures will be used in this study:

1) Architectures that create static embeddings. They create vector representations of words by using a large amount of input text and simultaneously create a vector space where the word representations are arranged. These representations take into account the general semantics of the words, but they do not consider the different senses that a single word can have (such as, for example, those produced by the algorithm called *Word2Vec)*.

2) Architectures that create contextualized embeddings, created by language models that at an earlier stage (called pre-train) take as input a large amount of unlabeled text. Then at the stage where the representation is produced, the language model receives as input a sentence, and at that time, the model will produce a vector representation of the words in the precise context of the sentence provided as input. Thus, by changing the context of the sentence, the word's representation will also change, even considering the embedding of the same word.

5.6. Similarity measure

The similarity measure used in this study is called *cosine similarity*. Cosine similarity is a measure of similarity between two nonzero vectors, and it is defined by the cosine of the angle between the two vectors (thus, of the word representations). To calculate this similarity, we take into account the angle that lies between the two lines passing through the origin of the axes and also passing through the vectors representing the two words. The cosine function is used as normalization to simplify the comparisons; in fact, the positive values of the cosine of an angle are in a range between 0 and 1. The cosine of 0° will be maximum and thus equal to 1, and in this case, also the value of maximum similarity will be attained: in fact, an angle of 0° means that the words under consideration are the same word, so the vectors will correspond and have the same direction and position, moreover no angle other than 0 will stand between the two.

Supported by a visual approach, it is possible to show that if a vector (or embedding) called $v1$ is represented in a space, vectors that represent words semantically related to $v1$, are also spatially arranged as having a position that is close to $v1$, and the vectors of the words belonging to different semantic fields are represented in a distant area of the vector space (Jurafsky, Martin 2014).

The space where vectors are represented is distributed and continuous, and for this reason, it is possible to talk concretely about the distance between two vectors as a similarity measure; this similarity can be computed by considering the distance of the vectors in the space they are represented into (Pilehvar, Camacho-Collados 2020).

5.7. Related works

Early approaches to the study of vector representations were based on collecting statistics on words, often referring to their occurrences, co-occurrences, and frequencies. In the early 2000s, *Deep Learning* techniques brought novelty to the field. In 2013, Mikolov *et al*. published Word2Vec[5]. Word2Vec is a group of models (or algorithms) used to

[5] Word2Vec is an algorithm taking as input a large corpus containing text and giving

produce embeddings. These models usually consist of two layers of neural networks trained through the input data to capture the relationships between words and their linguistic context. Vectors corresponding to words are placed in the vectorial space where, by design, vectors representing semantically similar words will occupy the space adjacent areas (Mikolov *et al*. 2013). However, using embeddings created with Word2Vec, it is impossible to seize the difference between the distinct senses that a single word may have.

A big step forward was made with the *Recurrent Neural Networks* (RNNs). RNNs are a type of architecture that receives data sequentially, processes the data received at each timestep, and maintains a state that contains the information processed up to that point. However, generic RNNs were complicated to use in real-life applications, mainly due to the enormous gradient computations that can become nearly impossible when calculating the partial derivative during the backpropagantion[6] (Jurafsky, Martin 2014). In order to find a solution to this problem, a new architecture, called Long Short-Term Memory (LSTM) was introduced in 1997 by Hochreiter and Schmidhuber, which by basing itself on RNNs, and thanks to a gating mechanism, it succeeded in overcoming the problem that occurred during backpropagation (Hochreiter, Schmidhuber 1997). In addition, LSTMs manage to handle the context of the sentence by not storing information deemed unnecessary in memory, thus saving only information likely to be needed in the future. In 2018 Peters *et al*., released ELMo (Embeddings from Language Models), an architecture based on LSTM, which allows the creation of contextualized embeddings.

However, the big breakthrough came in 2017, when LSTMs were ousted from being the state-of-the-art architecture for several language modeling tasks, and were then replaced by a class of architectures called *Transformer*. The first Transformer architecture was published

as output a vector space, usually having hundreds of dimensions, where each word belonging to the given input corpus is associated with a corresponding vector inside this space (Mikolov *et al*.).

6 Backpropagation is a machine learning algorithm used to calculate the gradient of the function that estimates the errors made by the neural network concerning the weights of the network itself. Backpropagation stands for "backward propagation of errors", in fact, the computation of the gradient is carried out in the opposite direction from the flow of the neural network (Goodfellow *et al*. 2016).

by Vaswani *et al*. in 2017, as a new approach based solely on the *Attention Mechanism*[7] and no longer on heavy architectures such as RNNs.

The generic Transformer proposed in this paper has an encoder-decoder structure. The input that both the encoder and decoder receive at first is a word embedding, followed by positional embedding[8]. Both the encoder and the decoder are composed of several stacks of layers; within these layers, computational operations take place, including the multiplication between matrices called Self-attention (Vaswani *et al*. 2017).

Nowadays, despite "Transformer" generally denoting a wide range of language models that may still differ significantly from each other, the similarity they all share is the use of the attention mechanism as the main structure. The differences usually lie in using different tasks to train the model during the training phase, or even in the number of parameters or in the structure of the model.

In 2018, Devlin *et al*. published a paper titled "BERT: Bidirectional Encoder Representation from Transformers", where they introduced their language model, which is also a transformer since it is based only on attention mechanism (Devlin *et al*. 2018). As the name itself suggests, BERT is a Bidirectional Encoder, which means that it does not come with an encoder-decoder structure like Vaswani *et al.*'s Transformer, but only with the encoding part; and it is bidirectional, which means it processes the sentences both from left to right and right to left (Devlin *et al*. 2018).

5.8. Empirical data: static embeddings and WordNet

The tests here presented are performed by calculating the distance

---

7 The Attention Mechanism consists of multiplications among the matrices of the vector that the model is considering at that moment, by the matrix of all other words is included in the sentence. This way, the resulting vector will contain information about the individual word but also about its context. The multiplication between matrices is followed by a scoring function (which is usually a *softmax* function) that normalizes the result and makes it a number between 0 and 1. So that it keeps intact the values of the words the model should focus on, and drown-out irrelevant words (Alammar 2018).

8 The positional embedding is the vector representation of the position of each word in a sentence. It is usually concatenated with the word embedding of each word (Devlin *et al*. 2018).

between pre-trained embeddings based on morphemes and not whole words. The embeddings used were created by Lee *et al*. in 2018. Given the structure of the Korean language, where even the single syllable (or morpheme) may have its own semantic relevance, it is beneficial to use these vectors. These embeddings were created by expanding the skip-gram model (of Word2Vec) to define each word vector as the sum of the vectors of the morphemes composing the words. The vectors were trained from a news corpus based on Naver's news (Lee *et al*. 2018).

Since the context where the word appears is not part of the vector representation, a full sentence will not be needed to analyze the similarity between the words; in fact, computing the cosine similarity between the vectors of the two words will suffice. Moreover, the representation of a large number of words is already available, so it will also be possible to plot them and find the most similar word list for any given word.

Pair: *saram- ingan* 사람 -인간	Similarity: 0,3506
10 most similar words	
saram 사람 (Noun)	*saram-gwa, jaga, jeolmeun-i, ja-reul, simin, nom, nugunga, pyeongsaeng, inmul* 사람과, 자가, 젊은이, 자를, 시민, 놈, 누군가, 평생, 인물
ingan 인간 (Noun)	*robot, inryu, joneomseong, jayeon, saramgwa,bonseong, jadonghwa, yokmang, saengmyeongche, sanghojakyong* 로봇, 인류, 존엄성, 자연, 사람과, 본성, 자동화, 욕망, 생명체, 상호작용

Tab. 5.1. Example showing similarity in the synonym pair *saram - ingan* 사람 – 인간 (person, human being).

The similarity between the pair *saram - ingan* 사람 – 인간 (person, human being)[9], shown in Table 5.1., is relatively low, to the extent that *ingan* 인간 is not in the list of the ten most similar words to *saram* 사람 and vice versa.

[9] A comma separating two translations means both are suitable for both the Korean words given. In case of different meaning of two Korean words, "and" will be used, instead, with a correspondence 1:1.

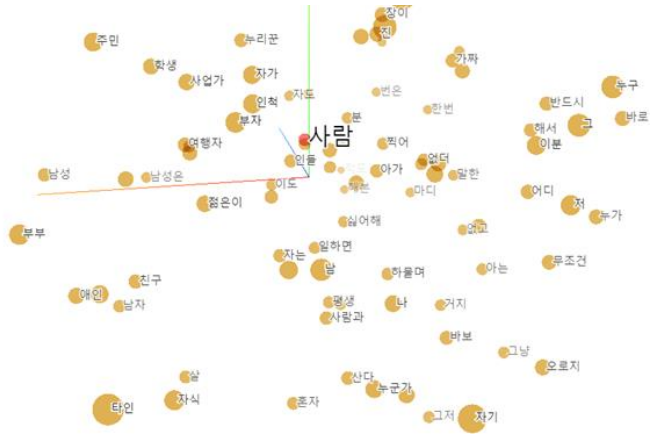

Figure 5.1. A three-dimensional projection of the words whose embedding is most similar to *saram* 사람 (person).

Pair: *nai – yeonnyeong* 나이 **-** 연령	Similarity: **0,6088**
10 most similar words	
nai 나이 (Noun)	*yeonnyeong, sal-e, seongbyeol, seoreun, sal-i, eorin, jalmeun, harabeoji, sein* 연령, 살의, 성별, 서른, 살이, 어린, 젊은, 할아버지, 세인
yeonnyeong 연령 (Noun)	*nai, sei, seongbyeol, gidaesumyeong, seongogwon, goryeongja, pyeong-gyunyeonryeong, yubyeongryul, gagu, sa-mangryul* 나이, 세이, 성별, 기대수명, 선거권, 고령자, 평균연령, 유병률, 가구, 사망률

Tab. 5.2. Example showing similarity in the pair of synonym *nai - yeonryeong* 나이 – 연령(age, years).

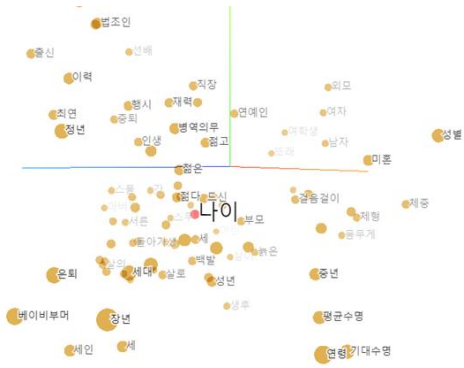

Fig. 5.2. Tridimensional projection of some of the words whose embedding is most similar to the one of *nai* 나이.

In Table 5.2., the similarity between the pair *nai - yeonryeong* 나이 - 연령 (age, years) is higher than the one in Table 5.1., noting that *nai* 나이 appears as the first word most similar to *yeonryong* 연령.

So, in this case, we will need a counter-test to see if there is a reason that can explain the low similarity of the first pair and the higher similarity of the second pair.

Pair	Similarity
chimdae – gaeguri **침대 - 개구리** (Noun)	**0.2705**
gangaji – goyangi **강아지 – 고양이** (Noun)	**0.7257**

Tab. 5.3. Example showing similarity values between the words *chimdae - gaeguri* 침대 - 개구리 (bed and frog) and *gangaji - goyangi* 강아지 - 고양이 (dog and cat).

Among the word pairs tested, in Table 5.3., as it might be expected, very low similarity for the pair *chimdae - gaeguri* 침대 - 개구리 (bed and frog) was found because the words do not share a common semantic field. The similarity between the pair *gangaji - goyangi* 강아지 - 고양이 (dog and cat) is also rather high. In the dog-cat pair, the two words share the semantic field and usually appear in very similar contexts, so it is interesting that they happen to be more similar than any other two words that are actually synonyms.

At this point, a question may arise: can we say that two words are synonyms just because the distance among the representations is low?

Apparently not, because the distance can be small even between two related words, such as "cat" and "dog". Therefore, to study synonymy, we need another resource that prevents us from incorrectly asserting that "cat" and "dog" are synonyms. This resource is WordNet, specifically its Korean version.

The Korean WordNet (KWN) was created by Semantic Web Research Center at KAIST (Jiseong Kim, Ingeun Lee, Key-Sun Choi), built using CoreNet and dictionaries from various domains. KWN includes lemmas, definitions, examples for synsets, and case frames for predicates. In a WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (called synsets), each expressing a distinct concept; KWN contains 9714 synsets. The synsets are connected by conceptual-semantic and lexical relationships. A synset includes a short definition (called "gloss") and, in most cases, one or more short examples illustrating the use of the synset's lemmas.

Thus, Korean WordNet was employed to check whether the two words with very high similarity values belonged to the same synset; if so, they would be synonyms; if not, the two words would only be related. This in theory, yet due to the lesser synsets included in the KWN (9714 synsets, compared to 117000 in the English WordNet), a synonym may also be missing, so a dictionary will be useful to manually check whether the two words share a meaning.

Pair *nai – yeonryeong* **나이 – 연령**	
Total senses for *nai* 나이	7
Total senses for *yeonryeong* 연령	2
Intersection size	2
Intersecting synsets	*gi, nonyeon, yeonryun, hae, nyeon, nai, yeonryeong, yeondae, yeongi, na, chunchu* (Noun) eld, age, 기, 노년, 연륜, 해, 년, 나이, 연령, 연대, 연기, 나, 춘추
	na, yeonryeong, nai, chunchu, yeonryun (Noun) age, 나, 연령, 나이, 춘추, 연륜

Tab. 5.4. Example showing synsets intersection (2) between the pair *nai - yeonryeong* 나이 - 연령 (age, years).

Pair *gangaji - goyangi* **강아지 – 고양이**	
Total senses for *gangaji* 강아지	2
Total senses for *goyangi* 고양이	4
Intersection size	0
Intersecting synsets	/

Tab. 5.5. In this example, we show the synsets intersection between the pair *gangaji - goyangi* 강아지 - 고양이 (dog and cat), and they turn out to be 0.

Pair *saram - ingan* **사람 – 인간**	
Total senses for *saram* 사람	44
Total of senses for *ingan* 인간	13
Intersection size	12
Intersecting synsets	
Noun - man, adult_male; *sanae, daejangbu, sinsa, jeongbu, inryu, nama, namseong, saram, namja, sanai, ingan* 사내, 대장부, 신사, 정부, 인류, 남아, 남성, 사람, 남자, 사나이, 인간	
Noun - individual, person, mortal, someone, soul, somebody; *in, ingan, inmul, saram, inryu* 인, 인간, 인물, 사람, 인류	
Noun - man, military_man, serviceman, military_personnel; *inryu, saram, jeonsa, byeongsa, gunsa, muin, ingan, musa, gunin* 인류, 사람, 전사, 병사, 군사, 무인, 인간, 무사, 군인	
Noun - Stiff; *inryu, saram, in, inmul, son, eoron, ingan, seongin* 인류, 사람, 인, 인물, 손, 어른, 인간, 성인	
Noun - public, world, populace; *maeseu, gongjung, seomin, saram, sein, sahoe, ingan, daejung, minjung* 매스, 공중, 서민, 사람, 세인, 사회, 인간, 대중, 민중	
Noun - valet, gentleman's_gentleman, gentleman, man, valet_de_chambre; *ingan, saram, inryu* 인간, 사람, 인류	
Noun - Man; *sanae, daejangbu, jeongbu, inryu, namseong, saram, namja, in, sanai, inmul, ingan* 사내, 대장부, 정부, 인류, 남성, 사람, 남자, 인, 사나이, 인물, 인간	
Noun - Person; *pyeon, got, inryu, saram, in, inmul, cheuk, ingan* 편, 것, 인류, 사람, 인, 인물, 측, 인간	
Noun - Man; *sanae, daejangbu, jeongbu, inryu, namseong, saram, namja, in, sanai, inmul, ingan* 사내, 대장부, 정부, 인류, 남성, 사람, 남자, 인, 사나이, 인물, 인간	
Noun - Man; *in, ingan, inmul, saram, inryu* 인, 인간, 인물, 사람, 인류	
Noun - homophile, homo, homosexual, gay; *ingan, homo, saram, inryu* 인간, 호모, 사람, 인류	

Noun - Man; *ingan, saram, inryu* 인간, 사람, 인류

Tab. 5.6. Example showing synsets intersection (12) between the pair *saram - ingan* 사람 - 인간 (person, human being).

So, running these queries on KWN returned the following results: each word pair searched showed the number of synsets (therefore different meanings) each word has, the number of meanings shared by the word pair (shown by the "intersection size" metric) and the list of shared meanings for the pair.

Nonetheless, given the small size of the Korean WordNet, it is complicated to use KWN to search extensively. Therefore, it was deemed necessary trying to use contextualized embeddings.

5.9. Empirical data: contextualized embeddings

To perform the tests that take into account the sentence context, a pre-trained[10] version of BERT to produce word embeddings was used to generate the word embedding, and then we computed the similarity measure of the native-Korean word and its Sino- Korean synonym. Since the embeddings generated with BERT are context-aware, they can be defined as the representation of the word within the sentence context.

We compared embeddings of native-Korean and Sino-Korean word pairs in different sentences and within same sentences (that only differed in the native-Korean or Sino-Korean synonym). We also compared the similarity of these synonym pairs with the similarity values of semantically related words, such as dog and cat, in two different contexts and within the same context. In the tests for same-sentence synonyms, the vector representations of the two words were very similar.

The Korean sentences were randomly selected from a news corpus. Where possible, we tried to ensure that the words whose similarity

[10] That is, for this study, we did not train the language model, but instead used the representations it produced.

was computed were connected to the same grammatical particle – defining their syntactic function in the sentence – in order to have them differ precisely only in the synonym used.

Pair: ***jari*** **–** ***jwaseok*** 자리 – 좌석	Similarity: 0,7807
sentence1 = *iddae chunseong-i* ***jari-eseo*** *beolddeok irreonaseo marhayeotta* "이때 춘성이 **자리에서** 벌떡 일어나서 말하였다." target_word1 = *jari-eseo* 자리에서 sentence2 = *iddae chunseong-i* ***jwaseok-eseo*** *beolddeok irreonaseo marhayeotta* "이때 춘성이 **좌석**에서 벌떡 일어나서 말하였다." target_word2 = *jwaseok-eseo* 좌석에서	
Pair: ***saram*** **–** ***ingan*** 사람 – 인간	Similarity: 0,8152
sentence1 = *geudeul-eul ilsijeok-euro* ***ingan-e*** *moseup-euro baggueojul mabeop iss-euni ganeunghan irieotta* "그들을 일시적으로 **인간**의 모습으로 바꾸어줄 마법이 있으니 가능한 일이었다." target_word1 = *ingan-e* 인간의 sentence2 = *geudeul-eul ilsijeok-euro* ***saram-e*** *moseup-euro baggueojul mabeop iss-euni ganeunghan irieotta* "그들을 일시적으로 **사람**의 모습으로 바꾸어줄 마법이 있으니 가능한 일이었다." target_word2 = *saram-e* 사람의	

Tab. 5.7. Two examples of similarity values among the synonym pairs *jari - jwasok* 자리 - 좌석 (seat) (pair 1) and *saram - ingan* 사람 - 인간 (person, human being) (pair 2).

In both cases shown in Table 5.7., it's worth noting a reasonably high similarity result, considering that 1 is the maximum value for this metric.

To check whether there was a pattern behind the positive results of the examples in Table 5.7., we performed the same similarity measure on another example with the pair *ingan - saram* 인간 - 사람 (human being, person). In the previous example, this pair showed high similarity between the words, so in the new example, the same pair of words were analyzed in different sentences (thus, different contexts). In this case, we expected a slightly lower similarity but still high enough to affirm that the words are synonyms.

Pair: *saram - ingan* 사람 – 인간	Similarity: 0,4476
sentence1 = *geudeul-eul ilsijeok-euro* ***ingan-e*** *moseup-euro baggueojul mabeop iss-euni ganeunghan irieotta*	

"그들을 일시적으로 **인간의** 모습으로 바꾸어줄 마법이 있으니 가능한 일이었다."
target_word1 = *ingan-e* 인간의
sentence2 = *hajiman **saram-e** insaeng-eseoe bomeun dasi doraoji anneunda*
"하지만 **사람의** 인생에서의 봄은 다시 돌아오지 않는다."
target_word2 = *saram-e* 사람의

Tab. 5.8. Similarity value in the synonym pair *saram - ingan* 사람 - 인간 (person, human being) in the context of two different sentences.

Although we expected a lower similarity value than the one shown by the example in Table 5.7., the result of the example in Table 5.8. is still a bit on the low side.

So, to better understand the situation, a test calculating the similarity between two words that have nothing in common – neither meaning nor context – was performed. This example can also be helpful in understanding the meaning of possible similarity values.

Pair: *chimdae- gaeguri*침대 – 개구리	Similarity: 0, 2584
sentence1 = *maejang tueoreul hal sigan-i eopseodo jip-e saram-i eopseodo **chimdae-reul** guiphal su itneun geosida*	
"매장 투어를 할 시간이 없어도 집에 사람이 없어도 **침대를** 구입할 수 있는 것이다."	
target_word1 = *chimdae-reul* 침대를	
sentence2 = *geu nal bam-e **gaeguri-e** boksilboksilhan dwitmok-eul seulseul sseudadeumgo itjanigga sinabeuro maekju saenggak-i nadda*	
"그 날 밤에 **개구리의** 복실복실한 뒷목을 슬슬 쓰다듬고 있자니까 시나브로 맥주 생각이 났다."	
target_word2 = 개구리의	

Tab. 5.9. Similarity value computed for the pair *chimdae - gaeguri* 침대 - 개구리 (bed and frog) in two different sentences.

The similarity value shown in Table 5.9. was expected to be very low due to bed and frog having no semantic field in common. Considering the similarity value between "bed" and "frog" in different sentences, it is interesting to note that the pair *saram - ingan* 사람 - 인간, when appearing in different sentences, still has a higher similarity value than "bed" – "frog".

Pair: *saram – saram* 사람 –사람	Similarity: 0,4779
sentence1 = *jultagineun iran **saram-deule** gwansim-eul gajang mani badatdeon nol-i gongyeonida* "줄타기는 이란 **사람들의** 관심을 가장 많이 받았던 놀이 공연이다." target_word1 = *saram-deule*사람들의 sentence2 = *hajiman **saram-e** insaeng-eseoe bomeun dasi doraoji anneunda* "하지만 **사람의** 인생에서의 봄은 다시 돌아오지 않는다." target_word2 = *saram-e* 사람의	
Pair: *gangaji – goyangi* 강아지–고양이	Similarity: 0,4789
sentence1 = *na-neun geu deok-e jadongcha geokjeong opsi jip mit-eseo **gangaji-reul** sanchaeksigineun hosa-reul nurigo idda* "나는 그 덕에 자동차 걱정 없이 집 밑에서 **강아지를** 산책시키는 호사를 누리고 있다." target_word1 = *gangaji-reul*강아지를 sentence2 = *baro wiheomhan sanghwang-eseo agi **goyangi-reul** gujohasin bun-deulijiyo* "바로 위험한 상황에서 아기 **고양이를** 구조하신 분들이지요." target_word2 = *goyangi-reul* 고양이를	
Pair: *gangaji - goyangi*강아지–고양이	Similarity: 0,9066
sentence1 = *oneul jeonyeok-e na-neun **gangaji-ege** meoki-reul jueodda* "오늘 저녁에 나는 **강아지에게** 먹이를 주었다." target_word1 = *gangaji-ege*강아지에게 sentence2 = *oneul jeonyeok-e na-neun **goyangi-ege** meoki-reul jueodda* "오늘 저녁에 나는 **고양이에게** 먹이를 주었다." target_word2 = *goyangi-ege* 고양이에게	

Tab. 5.10. Similarity value computed for the pair *saram - saram* 사람 - 사람 (per.son – person) in two different sentences (pair 1), then for the pair *gangaji - goyangi* 강아지 - 고양이 (dog and cat) in different sentences (pair 2) and in the same sentence (pair 3).

The three examples shown in Table 5.10. are handy for the purpose of the study: in the first case, it can be noted how even the same word (person, 사람), when in a different context, can be not so similar to itself. By comparing the similarity of the first pair with the similarity of the second pair, we can see that the similarity between "person" and "person" in a different context turns out to be almost the same as the similarity between "dog" and "cat" (which are only semantically related words) in a different context. However, the last example has the most interesting results: "dog" and "cat" are very similar when

appearing in the same sentence, even more similar than the pair *ingan - saram* 인간 - 사람 as in the example in Table 5.7.

What do these similarity values entail? The first thing that may come to mind is that there may be some errors with the similarity measure or with the vectors BERT generated, but that does not seem to be the case. The problem may also seem related to the computation of similarity between contextualized embeddings, for if the context where the words appear is too different, also the vector representation of the words will differ. If this is true, then why do two allegedly synonymic words appear in such a different context? A convincing explanation seems to be that, in most cases, native-Korean words can be used in a broader range of contexts than Sino-Korean words. Thus, we cannot look for synonymy by comparing the contextualized vector representation with different polysemy values.

5.10. More data

In this section, we take into account the similarity measure between the word pairs shown in the column "pair" (where the first word is the Sino-Korean one, and the second is the Native-Korean one). For each word pair, we consider the cosine similarity computed using vector representations from different encoding techniques (transformer-based language models and the static embeddings in the last column): in fact, we aim to inquire if using embeddings from different language models it is possible to get different results because that would mean the language model's objective function[11] and the data used during the train play an essential role in defining synonymy-like relationships between vectors.

The language models exploited in this study are BERT base trained on KLUE tasks, BERT base for Korean language, RoBERTa base, ELECTRA base for Korean and XLM RoBERTa base. The main difference between the two BERT versions is the train data, while RoBERTa is an

[11] The Objective Function is the task used to pre-train the language model. One of the most common functions is Masked Language Modeling, where some of the input tokens are replaced with [MASK], and then the model is supposed to reconstruct the original tokens.

optimized language model, XLM is a multilingual language model, and ELECTRA has a different objective function. An interesting case is the one from ELECTRA, where instead of masking the input, the pre-train is carried out by replacing some tokens with plausible alternatives sampled from a small generator network, and a discriminative model is trained to predict whether each token in the corrupted input was replaced by a generator sample or not (Clark *et al*. 2020).

In the last column of Table 5.11., "word_vectors", we find some values equal to zero, it means that the computation wasn't carried out properly due to the lack of the vector for at least one of the words of the pair.

Pair	Pair	Klue/bert-base	Kykim/bert-kor-base	Klue/roberta-base	Kykim/electra-l	Xlm-roberta-ba	Word_vectors
jutaek - jip	('주택', '집')	0.75	0.74	1	0.47	0.99	0.64
gagyeok - gapt	('가격', '값')	0.73	0.65	1	0.56	0.99	0.61
biyong - gapt	('비용', '값')	0.73	0.71	1	0.79	0.99	0.35
sayong - sseuim	('사용', '쓰임')	0.69	0.66	0.21	0.65	0.99	0
jangso - de	('장소', '대')	0.42	0.47	0.98	0.27	0.98	0.18
dobo - balgeoleum	('도보', '발걸음')	0.74	0.56	0.83	0.8	0.99	0.23
somang - baram	('소망', '바람')	0.63	0.61	0.96	0.81	0.99	0.29
uido - baram	('의도', '바람')	0.6	0.58	0.99	0.7	0.99	0.24
huimang - baram	('희망', '바람')	0.65	0.5	0.99	0.83	0.99	0.2
seonhohada - deo johahad	('선호하다', '더 좋아하다')	0.76	0.71	0.63	0.96	0.99	0
siksahada - meokda	('식사하다', '먹다')	0.75	0.4	0.81	0.66	0.99	0
sumyeon - jam	('수면', '잠')	0.8	0.83	0.99	0.87	0.97	0.51
eumsik - meoki	('음식', '먹이')	0.64	0.75	1	0.59	0.99	0.37
saenggak - neuggim	('생각', '느낌')	0.77	0.81	0.54	0.83	0.99	0.48
uigyeon - neuggim	('의견', '느낌')	0.75	0.53	0.9	0.87	0.98	0.2
taedo - neuggim	('태도', '느낌')	0.73	0.55	1	0.87	0.98	0.28
gibun - neuggim	('기분', '느낌')	0.85	0.83	0.99	0.81	0.98	0.53
simri - maeum	('심리', '마음')	0.81	0.66	0.99	0.87	0.99	0.29
simjeong - maeum	('심정', '마음')	0.75	0.74	0.98	0.74	0.99	0.62
gamjeong - maeum	('감정', '마음')	0.8	0.74	0.94	0.86	0.99	0.37
saenggak - maeum	('생각', '마음')	0.7	0.76	0.57	0.84	0.99	0.55
yeonghyangryeok - ipgim	('영향력', '입김')	0.72	0.5	0.98	0.9	0.99	0.53
bunno - noyeoum	('분노', '노여움')	0.71	0.57	1	0.8	0.99	0.18
haengbok - gibbeun	('행복', '기쁨')	0.84	0.77	0.83	0.92	0.98	0.52
haengbok - jeulgeoum	('행복', '즐거움')	0.81	0.76	0.09	0.95	0.99	0.45
huirak - gibbum	('희락', '기쁨')	0.56	0.6	0.14	0.83	0.99	0
huirak - jeulgeoum	('희락', '즐거움')	0.54	0.54	0.59	0.79	0.99	0
aesu - seolum	('애수', '설움')	0.57	0.68	0.21	0.93	0.99	0
aesu - seulpeum	('애수', '슬픔')	0.53	0.62	0.19	0.61	0.99	0
aesang - seolum	('애상', '설움')	0.54	0.6	0.34	0.85	0.99	0
aesang - seulpeum	('애상', '슬픔')	0.55	0.5	0.32	0.6	0.99	0
simjang - gaseum	('심장', '가슴')	0.78	0.78	0.95	0.92	0.98	0.4
damryeok - bojjang	('담력', '보짱')	0.65	0.64	0.81	0.84	0.99	0
ingan - saram	('인간', '사람')	0.77	0.77	1	0.79	0.99	0.35
yeonryeong - nai	('연령', '나이')	0.86	0.68	1	0.59	0.99	0.61
jwaseok - jari	('자리', '좌석')	0.82	0.85	0.61	0.94	0.99	0.29

Tab. 5.11. This table shows the comparison between the similarity measure computed using vector representations from different language models.

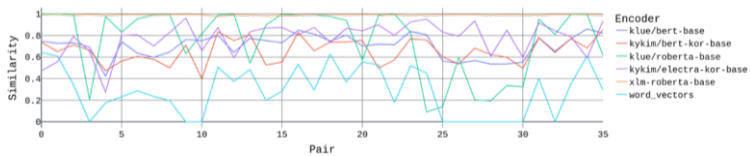

Fig. 5.3. This figure shows a plot of the values in Table 5.11.

Looking at Table 5.11. and Figure 5.3. is pretty clear that XLM RoBERTa is not the best model for this task, since it seems the model isn't really able to get the subtle differences between the word pairs; in fact, in Figure 5.3., the similarity values from XML RoBERTa plot a horizontal line. It could be due to the fact that XLM is a multilingual language model, so it performs better on other kinds of tasks.

Then, we can see the embeddings from the two BERT models give similarity measures very comparable, as we expected, because the model architecture is the same, while the pre-train data change. The static embeddings are those that give the lower similarity values, and it confirms the results shown in section 5.8.

RoBERTa shows some unexpected results: in fact, it returns some very low similarity values for pairs where the other language models return higher values, as in *sayong - sseumim* 사용- 쓰임, *haengbok - juelgoum* 행복-즐거움, *esang - seulpeum* 애상-슬픔.

So, it turns out that ELECTRA is the language model that returns more coherent similarity values, and it is probably due to the particular objective function that might allow the model to get a better generalization on synonymy.

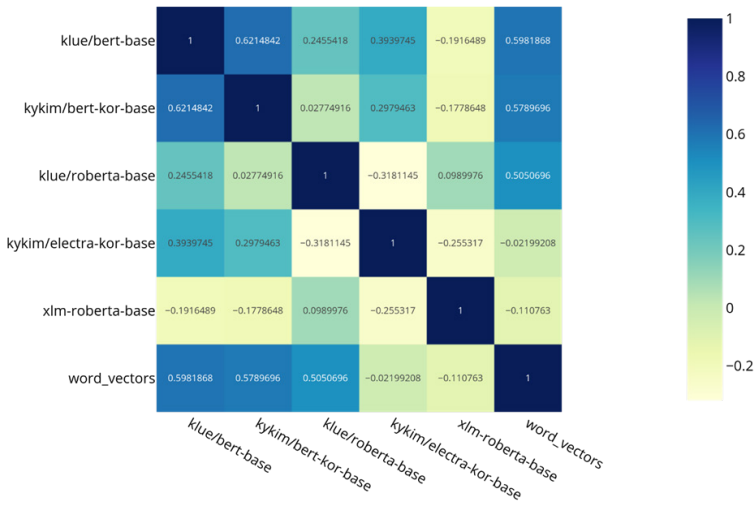

Fig. 5.4. Heat map showing Spearman correlation values on the data from Table 5.11.

To better understand the relationship among similarity measures on embeddings from different encoding techniques, we consider

Figure 5.4., which shows the Spearman correlation[12] on the data from Table 5.11. The correlation value shows how the similarity values computed using embeddings from a language model are similar to those from a different language model; we find values equal to 1 along the diagonal because those values express the correlation between each model and itself, but a correlation of 1 does not necessarily mean that the two sequences are the same. In fact, the correlation coefficient merely points out the relationship between two sequences: the higher the correlation, the easier it is to map each value in one of the two sequences to the corresponding value in the other one.

As we can see, Figure 5.4. confirms the previous analysis: the two BERT are pretty comparable, while ELECTRA and XLM are less comparable. It is interesting to notice that from Figure 5.4., we can see that the similarity values from the static embeddings are also quite close to those from the two BERT versions.

5.11. Linguistics perspective: number of meanings and senses, polysemy and homonymy

Polysemy is the feature of a word or phrase to have multiple semantically related meanings. A word is said to be polysemic when it is associated with two or more senses. It is the opposite of *monosemy*, that is a word form is associated with only one meaning, and also different from *homonymy*, where a single word form is associated with two or more unrelated meanings (Valera 2020).

A popular example in English is "bank": as a noun, the word can refer to the riverbank (or "levee" as "embankment") and also to "bank" as a financial institution. These two meanings of the word are called senses, that is, the meaning of the word from the context: if we find "money", "withdraw", and "branch" along with "bank", we are probably talking about the latter meaning, but if we find "river" or "dam" together with "bank", it is probable that we are talking about the former. In this case, since the senses of "bank" have very different

[12] Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables.

and unrelated meanings, we can say that it is a case of homonymy.

Senses	Definitions
ground ; land	The part of the earth that consists of soil or stones instead of water.
territory ; land	The territory of a nation.
town	A region or province.
plot ; land ; property	An inhabitable place where one can construct buildings or farms.
land ; soil	Soil or its characteristic.
ground ; place	An environment or basis in which one can enjoy his/her rights or work.
land	A society or world.

Tab. 5.12. The various meanings of the Korean word *ddang* 땅. Data were collected using the dictionary of the National Institute of Korean Language.

A different case is the one of *ddang* 땅 in Korean, which roughly means "land", even if the word has a lot of senses, since all belonging to the same semantic field (as shown in Table 5.12.) it is a clear case of polysemic word.

Polysemy is a property that some words have, and it is the ability of a word to have multiple related meanings. We distinguish polysemy from homonymy simply because homonymy is a mere linguistic coincidence. On the other hand, polysemy can be explained diachronically by observing word development; in fact, the use of pre-existing words in new contexts is a natural process of language changes. In natural language, the use of words with metaphorical, metonymical, synecdochic or other figurative meanings is widespread (Dancygier, Sweetser 2014; Valera 2020).

Therefore, it is reasonable that native-Korean words are more likely to be polysemous than Sino-Korean words, as they usually represent a specific concept and tend to have a less broad meaning. On the other hand, figurative meaning is much more likely to occur in basic concepts used in everyday language (Madigan 2015). Thus, while *toji* 토지 may be a synonym for *ddang* 땅 meant as "plot; land; property", it will not be a synonym for the other senses *ddang* of 땅.

5.12. Linguistics perspective: Light Verb Construction

The Light Verb Construction is a complex predicate construction where a light verb (i.e., a verb that does not carry solid semantics on its own) creates a complex predicate with the direct object of the verb, which in this case ceases to be a noun – and the object of the verb – but

acts as a carrier of "meaning" for the light verb (Kim 2016).

Korean has a very productive light verb construction process: it involves a verbal noun (usually a Sino-Korean noun) and the light verb *hada* 하다 (to do). The noun is the object of the verb 하다, so it should be marked with the object particle *reul* 를 as in *kongbu-reul hada* 공부를 하다 (to study), (where *kobgbu* 공부 is the noun) but the object marker can be omitted in specific contexts, so the noun is joined with the verb *hada* 하다 and the verbal phrase *gongbureul hada* 공부를 하다 becomes just one verb: *gongbu hada* 공부하다 "to study".

Interestingly, in Korean, the verb *hada* 하다 means "to do", but in this construction, it loses its meaning and works only as an auxiliary to turn the noun into a verb. Not all nouns can be part of this particular construct, even though it is more common for Sino-Korean nouns (usually consisting of two or three syllables) to be suitable as a noun for the Noun Phrase head of the light verb construction. The nouns taking part in this construction are called "verbal nouns" as they are syntactically nouns, but since the noun is also carrier of meaning, they also convey the information about the semantic role and subject of the light verb construction (Kim 2016).

piano-reul yeonseup jung-e jonhwaga wattda. 피아노를 연습 중에 전화가 왔다.
(piano-GEN practice while-at call-NOM come-PST-DECL)
«*in the midst of practicing piano, a call arrived*»

The "verb noun" *yeonseup* 연습 "practice" is a noun but behaves like a verb combining with the object *piano-reul* 피아노-를.

The interesting aspect of this construction is that it allowed the creation of new verbs with precise meanings from nouns carrying specific concepts. As Choo and Kwak (2008) state, Sino-Korean words are typically used in formal or literary contexts and to express abstract or complex ideas. Therefore, a large proportion of Korean verbs were created from a Sino-Korean noun and the verb *hada* 하다, and usually, these verbs tended to be monosemic. It is interesting to compare to native-Korean verbs, which do not have such a regular construction and are usually polysemic.

5.13. Conclusion

The purpose of this study was to point out that using static embeddings is not the right choice for inquiring about the synonymy of two words, as the embeddings are created in such a way that a single vector contains all senses of the word it conveys. Thus, the embedding of "mouse" contains a representation that is somewhere between the animal called "mouse" in English and the computer mouse. Therefore, if the words we are comparing do not share all senses, the similarity value will be lower than expected, because we are calculating the similarity of two representations that contain multiple meanings.

The use of contextualized embeddings can alleviate this problem because Transformer-based architectures use the sentence context where the word appears to create the embedding of the word. However, if we calculate the similarity between two words that appear in very different contexts, even if these two words share a common meaning, the similarity value might be quite low. In any case, contextualized embeddings can be used to study the synonymic relationship between two words, but to achieve that, it is recommendable to analyze the words within the same context. This means that if we want to analyze the synonymy of words such as "mouse" and "rat" we need to use two sentences involving the rodent animal, and not one sentence about rats and one about keyboards and mouse pads (which clearly refers to the computer-mouse sense).

A downside of the vector spaces used nowadays is that they are created in such a way that does not allow them to fully represent the semantic distribution of words globally. In fact, it is not particularly meaningful to compute the similarity between words that belong to different semantic fields, such as "frog" and "bed", as words that do not share a semantic field would be represented in portions of the vector space that are far apart, so in a way that it would not represent the semantic distribution of concepts in the language; therefore, this makes their similarity less meaningful to consider.

In the future, it would be interesting to have technologies that allow latent spaces to be created in a way for them to represent the semantic distribution of concepts, and in that case, the similarity between "frog" and "bed" will be close to 0, as we can assume.

Knowledge bases (such as WordNet) are valuable tools for

studying semantic relationships between words, as well as synonymy, for if we consider lemmas within a synset, we know that those lemmas are also synonyms for that particular sense. Unfortunately, the downside of using WordNet as a resource is that it is manually annotated, so the amount of data it can hold will always be limited by the fact that to be expanded, it will require people to annotate the data manually. We hope the Korean WordNet will be expanded because it currently contains only 9714 synsets, clearly not enough to study in depth the phenomenon of synonymy in a large portion of the Korean lexicon. Moreover, adding an example of the lemmas used in context for each synset would also be extremely helpful.

In conclusion, by performing queries on the Korean WordNet, it is confirmed that the pair consisting of a native-Korean word and a Sino-Korean word are synonyms for each synset where they intersect. This is due to the structure of WordNet, where the synset headwords must be synonyms by construction. Further studies could be done by analyzing the nuances of the synset lemmas or by working on the NLP task called Word In Context, but a larger and more complete dataset would be needed to carry out such studies.

Note on Romanization:

For this paper we chose to use Revised Romanization of Korean since it is the official romanization system in South Korea, developed by the National Academy of the Korean Language.

Vowel letters

Hangul	ㅏ	ㅐ	ㅑ	ㅒ	ㅓ	ㅔ	ㅕ	ㅖ	ㅗ	ㅘ	ㅙ	ㅚ	ㅛ	ㅜ	ㅝ	ㅞ	ㅟ	ㅠ	ㅡ	ㅢ	ㅣ
Romanization	*a*	*ae*	*ya*	*yae*	*eo*	*e*	*yeo*	*ye*	*o*	*wa*	*wae*	*oe*	*yo*	*u*	*wo*	*we*	*wi*	*yu*	*eu*	*ui*	*i*

Consonant letters

Hangul		ㄱ	ㄲ	ㄴ	ㄷ	ㄸ	ㄹ	ㅁ	ㅂ	ㅃ	ㅅ	ㅆ	ㅇ	ㅈ	ㅉ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ
Romanization	Initial	*g*	*kk*	*n*	*d*	*tt*	*r*	*m*	*b*	*pp*	*s*	*ss*	–	*j*	*jj*	*ch*	*k*	*t*	*p*	*h*
	Final	*k*	*k*		*t*	–	*l*		*p*	–	*t*	*t*	*ng*	*t*	–	*t*				*t*

Bibliography

AKITA KIMI, PARDESHI PRASHANT (2019), *Ideophones, Mimetics and Expressives*, Amesterdam-Philadelphia, John Benjamins Publishing Company.

ALAMMAR JAY (2018), *The Illustrated Transformer*, Blog Post, retrieved from <https://jalammar.github.io/illustrated-transformer/>.

BENGIO YOSHUA *et al*. (1993), "The Problem of Learning Long-Term Dependencies in Recurrent Networks", Conference Paper at the IEEE International Conference on Neural Networks, San Francisco.

BIRNER BETTY J. (2013), *Introduction to Pragmatics*, Chichester, Wiley-Blackwell.

BYON ANDREW SANGPIL (2017), *Modern Korean Grammar: A Practical Guide*, London, Taylor & Francis.

CHOO MIHO, KWAK HYE-YOUNG (2008), *Using Korean: A Guide to Contemporary Usage*, Cambridge, Cambridge University Press.

CLARK KEVIN *et al*. (2020), "ELECTRA: Pre-Training Text Encoders as Discriminators rather than Generators", Conference Paper at ICLR.

DANCYGIER BARBARA, SWEETSER EVE (2014), *Figurative Language*, Cambridge, Cambridge University Press.

DEVLIN JACOB *et al.* (2018), "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding", *arXiv preprint arXiv*, 1810.04805.

FIRTH JOHN RUPERT (1957), *Studies in Linguistic Analysis*, London, Wiley-Blackwell.

GOODFELLOW IAN J. *et al*. (2016), *Deep Learning*, Cambridge, Massachusetts, The MIT Press.

HOCHREITER SEPP, SCHMIDHUBER JÜRGEN (1997), "Long Short-Term Memory", *Neural Computation* 9.8, 1735-1780.

JURAFSKY DAN, MARTIN JAMES H. (2014), *Speech and Language Processing*, London, Pearson Custom Library.

KIM JONG-BOK (2016), *The Syntactic Structures of Korean: A Construction Grammar Perspective*, Cambridge, Cambridge University Press.

KIM MIN-JOO (2019), *The Syntax and Semantics of Noun Modifiers and the Theory of Universal Grammar: A Korean Perspective*, Cham, Springer

International Publishing, Studies in Natural Language and Linguistic Theory.

LEE DONGJUN *et al*. (2018), *Morpheme-Based Efficient Korean Word Embedding*, J. Korea Inf. Sci. Soc., Seoul.

LEE KI-MOON, RAMSEY ROBERT S. (2011), "Contemporary Korean", in Lee Ki-Moon, Robert S. Ramsey (eds.), *A History of the Korean Language*, Cambridge, Cambridge University Press, 287-305.

MADIGAN SEAN (2015), "Anaphora and Binding", in Lucien Brown, Yeon Jaehoon (eds.), *The Handbook of Korean Linguistics*, Chichester, John Wiley Sons, 137-154.

MIKOLOV TOMAS *et al*. (2013), "Distributed Representations of Words and Phrases and their Compositionality", Conference Paper at Advances in Neural Information Processing Systems.

PETERS MATTHEW E. *et al* (2018), "Deep Contextualized Word Representations", Conference Paper at NAACL.

PILEHVAR MOHAMMAD TAHER, CAMACHO-COLLADOS JOSE (2020), *Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning*, s.l., Morgan & Claypool Publishers, Synthesis Lectures on Human Language Technologies Series.

SEPP HOCHREITER (1998), "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6.2, 107-116.

SEUNGHO NAM (2015), "Lexical Semantics", in Lucien Brown, Yeon Jaehoon (eds.), *The Handbook of Korean Linguistics*, Chichester, John Wiley Sons, 155-178.

SHI CHUNG-KON (2015), "Word Formation", in Lucien Brown, Yeon Jaehoon (eds.), *The Handbook of Korean Linguistics*, Chichester, John Wiley Sons, 59-78.

SONG JAE JUNG (2005), *The Korean Language: Structure, Use and Context*, London, Routledge.

VALERA SALVADOR (2020), "Polysemy Versus Homonymy", *Oxford Research Encyclopedia of Linguistics*, internet ed.

VASWANI ASHISH *et al*. (2017), "Attention is All You Need", Conference Paper at Advances in Neural Information Processing Systems.

YEON JAEHOON, BROWN LUCIEN (2019), *Korean: A Comprehensive Grammar*, London, Taylor & Francis.

Collana Studi e Ricerche

Per informazioni sui volumi precedenti della collana, consultare il sito: www.editricesapienza.it | *For information on the previous volumes included in the series, please visit the following website: www.editricesapienza.it*

120. Multi-drug resistant Klebsiella pneumoniae strains circulating in hospital setting
Whole-genome sequencing and Bayesian phylogenetic analysis for outbreak investigations
Eleonora Cella
121. Agave negatively regulates YAP and TAZ transcriptionally and post-translationally in osteosarcoma cell lines
A promising strategy for Osteosarcoma treatment
Maria Ferraiuolo
122. Trigeminal Neuralgia
From clinical characteristics to pathophysiological mechanisms
Giulia Di Stefano
123. Le geometrie del Castello di Anet
Il 'pensiero' stereotomico di Philibert de l'Orme
Antonio Calandriello
124. Towards Recognizing New Semantic Concepts in New Visual Domains
Massimiliano Mancini
125. La distribuzione spaziale dei reperti come base per un'interpretazione dei livelli subappenninici di Coppa Nevigata (Manfredonia, FG) in termini di aree di attività
Enrico Lucci
126. Costruire, violare, placare: riti di fondazione, espiazione, dismissione tra fonti storiche e archeologia
Attestazioni a Roma e nel *Latium Vetus* dall'VIII a.C. al I d.C.
Silvia Stassi
127. Complexity of Social Phenomena
Measurements, Analysis, Representations and Synthesis
Leonardo Salvatore Alaimo
128. Etica ebraica e spirito del capitalismo in Werner Sombart
Ilaria Iannuzzi
129. Trauma Narratives in Italian and Transnational Women's Writing
edited by Tiziana de Rogatis and Katrin Wehling-Giorgi
130. Percorsi in Civiltà dell'Asia e dell'Africa II
Quaderni di studi dottorali alla Sapienza
a cura di Marina Miranda

Con il presente volume giunge al secondo tomo l'iniziativa editoriale inaugurata nel 2021, associata a un progetto precedente e volta a valorizzare e diffondere i risultati delle ricerche di giovani studiosi che stanno formandosi nell'ambito del Dottorato in Civiltà dell'Asia e dell'Africa, presso l'Università di Roma Sapienza. I saggi qui proposti, i cui autori sono iscritti al 36° e 35° ciclo, rispecchiano alcune delle principali specializzazioni del corso in questione e spaziano dalla letteratura sanscrita, cinese e giapponese alla linguistica coreana, dalla storia degli Studi orientali ad indagini etnografiche in Giordania. Di carattere multidisciplinare e basati su fonti in lingua originale, tali studi assumono particolare rilevanza in campo accademico, arricchendo i temi trattati con analisi innovative; allo stesso tempo, a un livello maggiormente divulgativo, essi contribuiscono a una più ampia comprensione delle culture asiatiche e medio-orientali per i diversi periodi e ambiti disciplinari considerati.

Marina Miranda è professoressa ordinaria di Storia della Cina contemporanea all'Università di Roma "Sapienza" e, presso lo stesso Ateneo, responsabile scientifico della sezione Asia Orientale del Dottorato in Civiltà dell'Asia e dell'Africa, di cui è stata Coordinatrice per due mandati, fino al 2018. È Presidente dell'Associazione Italiana per gli Studi Cinesi (AISC), Direttrice della Collana "Studi Orientali" (LibreriaUniversitaria.it) e membro del Comitato scientifico e redazionale di alcune note riviste. Oltre che di numerosi saggi, è autrice e curatrice dei seguenti volumi: *L'Identità Nazionale nel XXI Secolo in Cina, Giappone, Corea, Tibet e Taiwan* (2012); *La democrazia in Cina: le diverse formulazioni dagli anni '80 a oggi* (2013); *La Cina dopo il 2012* (2013); *Politica, società e cultura di una Cina in ascesa* (2016); *La Cina quarant'anni dopo Mao* (2017); *Ideologia e riforma politica in Cina: una democratizzazione elusa dagli anni Ottanta in poi* (2022).

ISBN 978-88-9377-260-0

9 788893 772600

www.editricesapienza.it