Unveiling Middle-Level Concepts through Frequency Trajectories and Peaks Analysis

Luigi Di Caro*
Dept. of Computer Science, University of Turin
Turin, Italy
dicaro@di.unito.it

Alice Ruggeri
Dept. of Computer Science, University of Turin
Turin, Italy
ruggeri@di.unito.it

ABSTRACT

Natural language development and learning begins with a few terms, often related to concepts associated to perception, in general neither too general nor too specific, gradually specializing and extending to abstract concepts. A *linguistic middle-level* can be considered as a lexical base on which a language oriented towards more complex communications develops in adulthood. Learning can extend from generic to specific terms (usually a child learns first the term "*dog*", then the term "*dachshund*"), or from specific to generic ("*dog*" → "*mammal*"). Thus, a middle-level includes all those terms that are halfway, in the taxonomic sense, between generic and specific concepts. In this paper we propose a computational approach to identify a middle-level relying on Wikipedia as input corpus and WordNet as lexical-semantic resource. An experimentation based on graded readings shows favourable results. The impact of this work could touch different fields, e.g., the automatic evaluation of the complexity of a text by computational approaches, rather than psychological and linguistic studies related to the language use.

CCS CONCEPTS

• **Computing methodologies** → **Cognitive science**; Lexical semantics;

KEYWORDS

Linguistic middle-level, Cognitive Aspects of Language, Language Development

ACM Reference Format:
Luigi Di Caro and Alice Ruggeri. 2019. Unveiling Middle-Level Concepts through Frequency Trajectories and Peaks Analysis. In *The 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19), April 8–12, 2019, Limassol, Cyprus.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3297280.3297383

*Corresponding author.

SAC '19, April 8–12, 2019, Limassol, Cyprus

ACM ISBN 978-1-4503-5933-7/19/04.
https://doi.org/10.1145/3297280.3297383

1 INTRODUCTION

Psycholinguistics (or psychology of language) can be defined as the study of psychological and neurobiological factors underlying acquisition, understanding and use of language in human beings.

In this context, in 1958, Professor Roger Brown published a paper on language development [1] based on a basic consideration: when we refer to a thing (which could be a concrete or abstract concept) we tend to use different names depending on contextual factors. Mr. Smith's dog is not only a "*dog*", but also a "*boxer*", a "*quadruped*", an "*animal*", and maybe it even has a proper name, like "*Prince*". Which of these words would a child use to identify him? Sometimes he would call him "*dog*", other times "*Prince*", less often he would call him "*animal*" or "*boxer*" and almost never would call him "*quadruped*".

So one wonders: what determines the name that a child uses to identify a certain thing? Or better yet, why does one prefer one name over another? One might think that short words are preferred due to the fact that children generally find it difficult to pronounce longer and more complex words. In this regard, it has been shown that the length of a word (in syllables) is inversely proportional to its frequency in the language. This principle based on the frequency-length of words works for the choice of the word "*dog*" to identify Mr. Smith's dog, but it is not valid in general. In fact, sometimes, following this criterion, it happens that the prediction of the chosen word is wrong. For example, the object called "*geranium*" can also be identified with the term "*flower*". In general, in the language, the word "*flower*" comes used much more frequently than the word "*geranium*" (and is also shorter), but if we consider a context in which we talk about geraniums, we can say that the word "*geranium*" is likely to be chosen more frequently compared to the word "*flower*" to refer to that concept. A short excerpt of the original Brown's paper is given.

> *Word counts of general usage are very roughly applicable to the prediction of what will be said when something is named. What we need is referent-name counts. We do not have them, but if we had them it is easy to see that they would improve our predictions.*

From the words of Professor R. Brown it emerges that the mere frequency of words in language is not indicative of how often a term is used to refer to a certain concept. It is simply an indication of how important that term is in the data collection. What we need to know is the frequency of *referent-names*, or the frequency of words associated with the concept to which one refers. Linking to the previous example, the referent-name would be the pair $< geranium, flower >$ and the frequency would represent the number of times the word "*flower*" is used to refer to the geranium object (the

referent). If one had knowledge of this type of frequency, concludes R. Brown, it would be possible to improve the prediction of the choice of the word used to express a certain concept.

The idea of this work arose precisely from this last point. In particular, we want to propose a computational solution that tries to identify a *linguistic middle-level*. By middle-level we mean a subset of the global lexicon of a language that identifies, roughly, a basic level of language. In fact, it is known that natural language learning starts from a few terms, often related to perceptual concepts, in general neither too general nor too specific, gradually specializing and extending to abstract concepts. The middle-level can be considered as the language of children, that is the basic linguistic base on which a language oriented towards more complex communications develops in adulthood. It can also be considered as a basic and simplified language for non-native speakers. Or still, that set of terms that allow us to describe every concept.

Learning can extend from a generic to a more specific term (usually a child learns first the term "*knife*", then the term "*cleaver*"), or from a more specific term to a more generic one (e.g., "*cutter*"). The middle-level therefore includes all those terms that are halfway, in the taxonomic sense, (hence the name middle-level) between the more generic concepts and the more specific concepts.

One of the most used resources in the world today is Wikipedia. Wikipedia is a well-known online encyclopedia with collaborative, multilingual and free content, born in 2001, supported and hosted by the Wikimedia Foundation, a non-profit US organization. Being an encyclopedia, by its nature, it contains words that are used to describe things. Specifically, such "things" approximately correspond to all words of a language. Making a simplification, it is as if Wikipedia were a tangible representation of all possible concepts expressed in language. Precisely because of this feature of completeness of the content, having available the texts that describe "all" the expressible concepts, if we calculate the frequency of the terms expressed, we obtain an aggregation of the *referent-name* frequencies. Taking up the previous example, calculating the frequency of the word "*flower*" is simultaneously calculating how often the word "*flower*" is used to describe all possible concepts (hence the sum of the frequencies of all possible pairs $< [referent], flower >$). This is the idea behind the search for middle-level concepts.

2 TRAJECTORIES AND PEAKS

As mentioned in the introduction, the middle-level does not correspond simply to the set of the most used terms of a language, but rather to those terms expressing all possible concepts through, again, that language.

Thus, it is necessary to think about the relationships of hypernymy and hyponymy between the concepts expressed by the terms, introducing the notions of *semantic trajectory* and *trajectory peak*. A semantic trajectory is simply a leaf-root path in a conceptual taxonomy (e.g., the WordNet taxonomy). A trajectory peak is a point in a semantic trajectory corresponding to a concept that presents a more frequent use than its direct hyponym and hypernym. This generally means that a term corresponding to a peak is a preferred term for describing a local range of a semantic trajectory.

According to these criteria, and by means of some basic NLP (Natural Language Processing) techniques, an algorithm has been studied and implemented to identify the set of peaks (within all semantic trajectories in WordNet) from the senses extracted over the 500 million terms (nouns and verbs) in Wikipedia.

3 DATA AND TOOLS

Before starting to illustrate the details of the proposed method, there are two important premises to make. The first premise is that the linguistic base on which the whole idea was developed is English. In particular, a Wikipedia dump in English (text content only, updated in November 2014) was used as input body. The Wikipedia pages have been indexed by Apache Lucene, an open source software designed for full-text indexing and searching. The reasons for choosing Wikipedia have been previously clarified. The second premise is that in the search for the middle-level the terms that correspond to names and verbs were the only considered.

The reasons for this choice are related to the nature of the grammar. Determiners, prepositions, conjunctions and pronouns represent types of words which are certainly very common in a language and are obviously indispensable to formulate sentences with a complete meaning. However, unlike nouns and verbs, they are not associated with concepts and therefore are not suitable to be included in a middle-level. On the other hand, adverbs and adjectives represent an exception which can in some way be associated with more or less abstract concepts. In spite of this, they are often secondary parts in terms of expressed meaning. In this paper, it was therefore decided not to take them into consideration.

The proposed approach can be viewed as a whole in two macro phases:

- the first phase consists of the analysis of the entire textual content of Wikipedia, from which we extract the relevant terms (nouns and verbs). At this stage, various Natural Language Processing techniques are used. Furthermore, the disambiguation of terms is particularly important in this context.
- the second phase consists of the search of the middle-level: a specific algorithm is applied based on the overall frequency of the terms in Wikipedia and on the above-mentioned concept of trajectory peak. This algorithm will be explained in detail in the next section.

For the Natural Language Processing pipeline, we made use of Stanford Core NLP (version 3.9.1), a java library offering an extensive list of functionalities such as Part-of-Speech Tagging. As an interface to WordNet, we used JWI (version 2.4.0), a library written in Java.

In the next section, the most relevant technical issues and aspects of the proposal will therefore be analyzed and discussed.

4 PROPOSED METHOD

The basic idea behind the search for middle-level terms is based on two key factors: semantic trajectories and trajectory peaks, previously described.

Consider any word that is a noun or a verb in Wikipedia. Depending on the context, it will have a corresponding synset in WordNet. This, in turn, will have hyponyms and hyperonyms, that is, it will be part of a certain part of the WordNet taxonomy. It is therefore easy to imagine WordNet as a huge set of leaf-root branches where the

leaves are the most specific synsets (i.e., those synsets that do not have hyponyms) and the roots are the most generic ones (i.e., those that do not have hyperonyms). Each of these leaf-root branches is as if it were an overall representation of a certain semantic sector, a sort of global semantic trajectory.

For example, consider the synset corresponding to the word "*book*": "*a written work or composition that has been published*". This has a long list of hyponyms, including e.g., "*songbook*". In fact, "*songbook*" is a specification of book ("*a book containing a collection of songs*"). The latter, in turn, has a hyponym sense: "*hymnal, hymnbook, hymnary*". In fact, hymnal is a type of songbook, and even more generally it is a type of book ("*a songbook containing a collection of hymns*"). Finally, hymnal has no hyponyms in its turn, so it consists of the leaf of the book-songbook-hymnal trajectory, and more generally consists of a leaf in WordNet. Returning to book and making the same path as the one just shown, but in the opposite verse, i.e., analyzing the hyperonymy relationships, it is possible to reconstruct also the upper part of the whole leaf-root branch. As mentioned above, this branch therefore represents a particular semantic trajectory. At this point, it is necessary to ask a question: which of the terms (or rather, the synsets) of this branch is used more commonly in the language in order to identify all the others? Or, which of these terms would a child use within the entire semantic sector?

The situation of the entire leaf-root branch in question is shown in Figure 1 (at the top there is the root, followed by the following hyponyms up to the leaf). Figure 2 shows the map of the sense frequencies of the same example of Figure 1.

A peak corresponds to a synset, within a leaf-root branch, whose overall frequency in Wikipedia is greater than its hyponym (if it has one) and its hyperonym (if it has one). Based on the concept of peak, therefore, it is possible to select those terms that turn out to be the most used in the language to describe a certain semantic trajectory, considering them therefore part of the middle-level.

It is important to consider the peaks, not just the frequency, precisely because the natural language develops both upwards (hyperonymy) and downwards (hyponymy). It would be, in fact, reductive to consider part of the middle-level only the term represented by the highest frequency, in this case "*book*", to the detriment of other more specific or more generic but equally important terms, such as object. On the other hand, it is also true that by doing so, a peak may result in a lower frequency than another non-peak synset within the same branch, such as for example with respect to "*publication*". In fact everything depends on the neighbors (hyponyms and hyperonyms): a peak, albeit at low frequencies, corresponds to the term that best identify, in addition to itself, even its direct hyponym and hyperonym (or a cascaded sequence of them). In essence, a peak is the most frequent term for describing a sub-branch of the entire leaf-root branch. The set of peaks, therefore, best represents the branch in its entirety.

At this point it is possible to implement an optimization: this consists in discarding those peaks that can be considered less relevant than others. Taking up the previous example, it is easy to notice how "*book*" is a peak with a much higher frequency than all the others. Moreover, in the right part of the graph one can see how the three peaks "*artifact*", "*object*" and "*entity*" are very close to each other, both as position (they are separated by a single sense

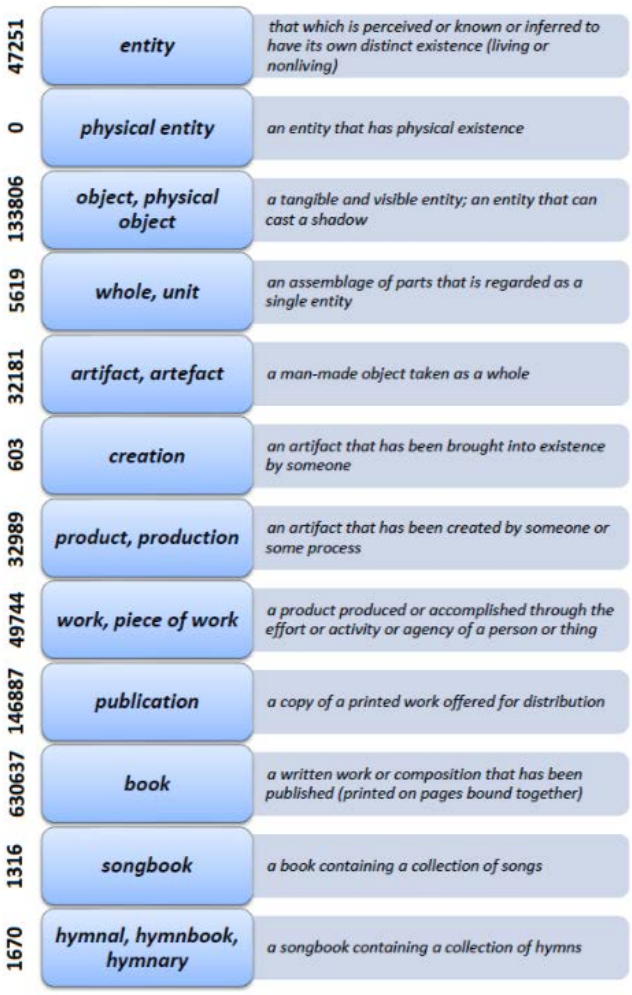

Figure 1: Leaf-root branch (semantic trajectory) for the leaf concept "*hymnal, hymnbook, hymnary*".

from each other) and as frequency (the frequencies are relatively similar to each other). In a case like this, the three peaks in question can be considered as a single group of peaks represented by the one with the highest frequency, in this case "*object*". The situation then becomes the following, illustrated in Figure 3.

Entering the implementation details, this optimization is carried out following the following list of operations:

- the average frequency deviation is calculated between each synset of the branch and its hypernym. Let name this value m;
- the mean square deviation of the deviations between each synset and the subsequent deviation from the mean deviation is calculated. Let this value be σ;
- for each peak previously detected, if the total deviation between adjacent peaks is greater than $(m + \sigma)$, then the peak is to be considered relevant, otherwise it is discarded.

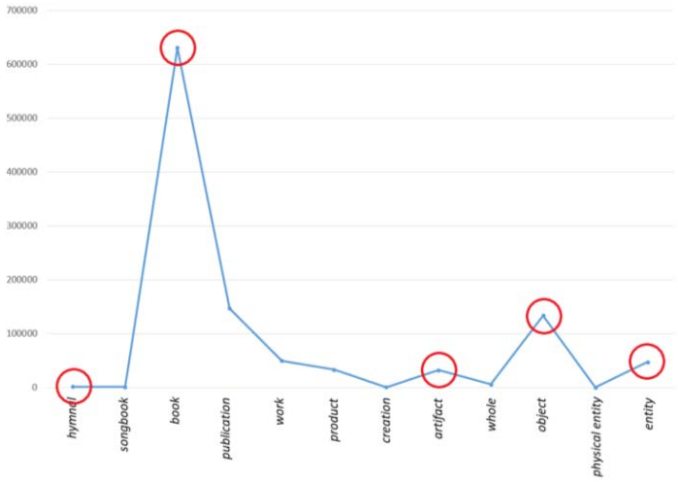

Figure 2: Semantic trajectory peaks for the leaf concept "*hymnal, hymnbook, hymnary*".

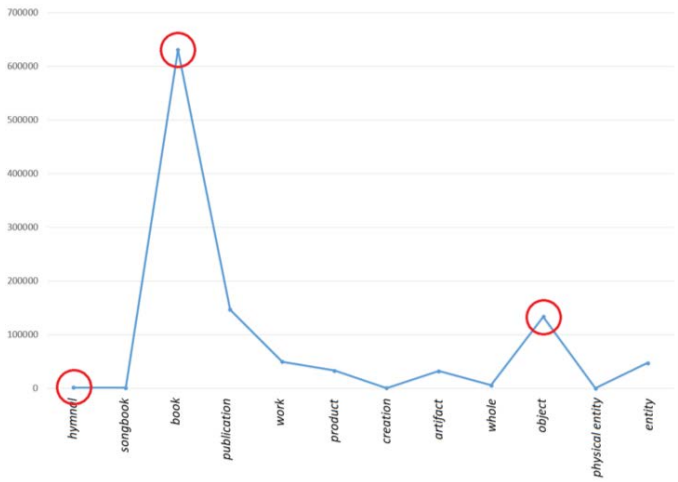

Figure 3: Semantic trajectory peaks for the leaf concept "*hymnal, hymnbook, hymnary*", after optimization.

The meaning behind this reasoning is as follows: M represents the average deviation between synsets. σ indicates instead the dispersion of the deviations, that is a value that represents how much the synsets differ from each other in a uniform way. The smaller the σ, the more the synsets are distributed evenly, the greater is σ, the more the synsets are distributed unevenly. The sum $(M + \sigma)$ is therefore a value that, taking into account how the data are distributed, indicates on average the deviation that actually exists between synsets. In light of this, two peaks that differ from each other by a value lower than $(M + \sigma)$ may be approximable by only one of the two.

By applying these operations to the previous example, we have $M = 147.918$ and $\sigma = 199.641$. The fact that σ is very large and even greater than M should not surprise, in fact by observing Figure 2 one can immediately notice a very dispersive distribution of the data: "*book*" is a very pronounced peak with respect to the frequency of the remaining synset that turns out to be more thickened. In practice, the synsets differ from each other on average M with deviation σ, i.e., 147.918 ± 199.641.

Considering therefore the peaks previously identified, it results that those corresponding to "*artifact*", "*object*" and "*entity*" deviate from one another by a value lower than $(M + \sigma)$, therefore among the three peaks the one with the highest frequency is chosen as representative, i.e., "*object*", while the remaining two are discarded.

At this point, the remaining peaks can be considered as components of the desired middle-level. But it is good to make a clarification: the WordNet synsets which are not peaks in a specific leaf-root branch could be peaks in other branches. Consider, for example, the semantic trajectory shown in Figure 4 for the leaf concept "*republication*". Applying to this new semantic trajectory the proposed method, "*publication*" turns out to be one of the peaks (hence, it is part of the middle-level). The same goes for "*artifact*" and "*entity*".

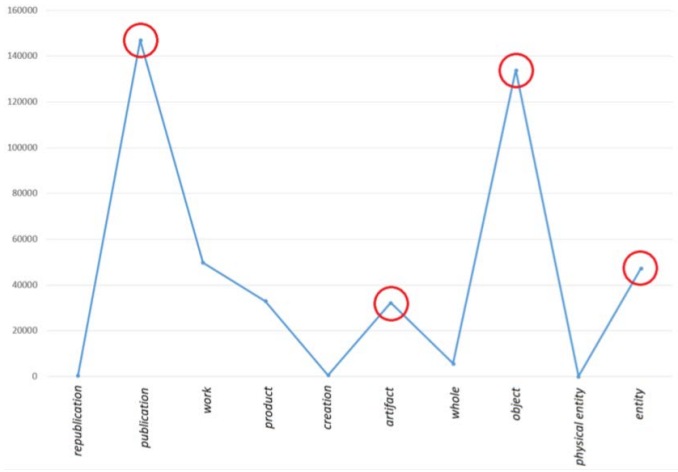

Figure 4: Semantic trajectory peaks for the leaf concept "*republication*".

Concluding the interpretation of the running examples, the meaning of trajectories / peaks is that while e.g., "*publication*" is not sufficiently important for book-related concepts such as "*songbook*" (see Figure 3), it is crucial to describe the leaf concept "*republication*" (see Figure 4).

Finally, a synset can be a peak in more than one semantic trajectory. Table 1 shows the distribution of the presence of synset peaks over WordNet root-leaf branches. The majority of the peaks are only discovered in a single semantic trajectory, while only 134 middle-level concepts (the sum of the last three rows of Table 1) appear in more than 100 trajectories. This analysis opens further research on cross-topic hierarchical structuring of the middle-level. Figure 5 shows an illustration of a leaf with more than one semantic trajectory. The central node (marked in black) can potentially be a middle-level concept within a single trajectory only, depending on the relative frequency value of its two direct hypernyms.

N. of root-leaf branches	N. of peaks
1	10128
[2,10]	3119
[11,100]	809
[101,1000]	115
[1001,10000]	16
More than 10000	3

Table 1: Presence of synset peaks in WordNet root-leaf branches.

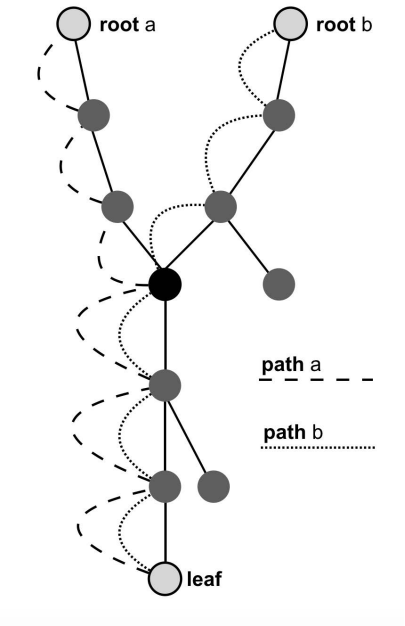

Figure 5: Case of multiple paths. The central node marked in black can be a middle-level concept in a single path only.

N. of sentences	76.532.404
N. of words	514.927.589
N. of nouns	305.436.881
N. of verbs	209.490.708

Table 2: Size of the used Wikipedia Dump.

5 RESULTS

In this section, we report the outcomes of the proposed method in terms of quantitative and qualitative measures.

5.1 Size of the problem

To give an idea of the size of the addressed problem, we report some data on the content of the Wikipedia dump used (see Table 2) and the subsequent development phases.

To each word corresponds a synset, obtained after a process of disambiguation (with the simple Lesk algorithm [7]). Keeping track of how often a certain synset appears in Wikipedia, the map of the sense frequencies is then created. Table 3 shows the WordNet synset frequencies, while Table 4 reports the number of leaves and semantic trajectories[1] and Table 5 the number and proportion of peaks with respect to the synsets.

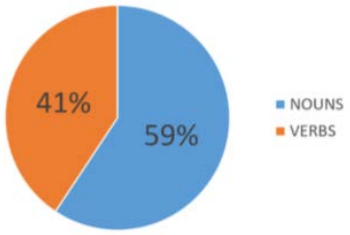

Figure 6: Words in Wikipedia.

N. of synsets	64.620
N. of noun synsets	53.191
N. of verb synsets	11.429

Table 3: Map of the sense frequencies.

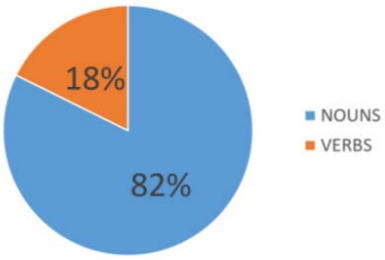

Figure 7: Synsets in Wikipedia.

N. of leaves	72.271
N. of semantic trajectories	84.932

Table 4: Number of leaves and leaf-root branches (i.e., semantic trajectories) in WordNet.

Figures 8 and 9 report a global and a zoomed-in view of the frequencies of the detected peaks.

From the graph it is immediately evident that there is a considerable quantity of peaks with a relatively low frequency. Going then to eliminate those peaks whose frequency is below a certain limit threshold, it is possible to make a significant reduction of the middle-level. The value to be used as a threshold can be chosen based on the frequency of words, considering "not relevant" those words in Wikipedia in less than 0.01% of cases (i.e., less than once in 10.000). Bearing in mind that the total number of words (counting

[1]Note that the number of trajectories is bigger than the number of leaves due to the existing multiple paths to the roots.

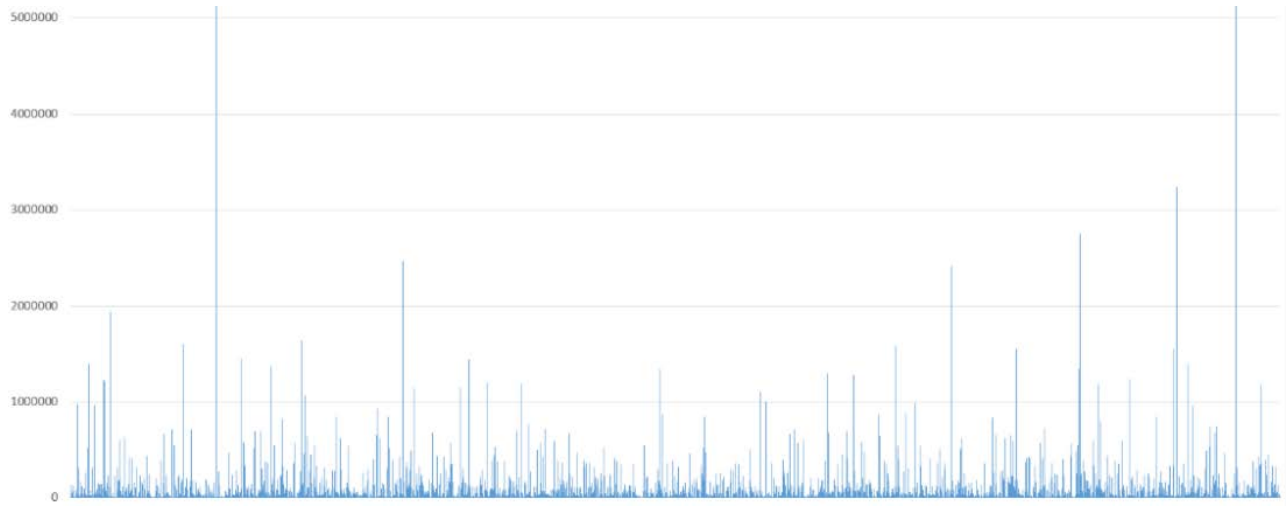

Figure 8: Global view of the identified peaks with their frequencies.

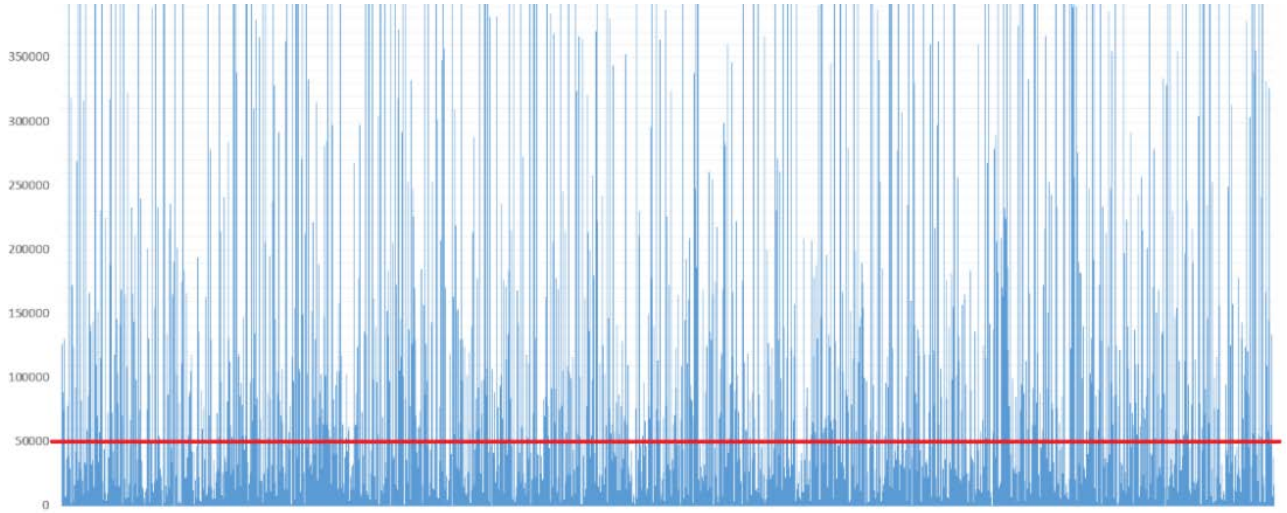

Figure 9: Zoomed-in view of the bottom part of the identified peaks of Figure 8. The red line indicates the applied threshold for the middle-level final extraction.

N. of peaks	14.190
Peaks / synsets ratio	21,96%

Table 5: Number of identified peaks and ratio with synsets.

only nouns and verbs) present in Wikipedia is just over 500 million, an estimate of the threshold value in question could be 50.000. The following graph consists of a zoom of the lower area of the previous graph and the red line corresponds to the chosen threshold.

The meaning behind this reasoning is as follows: a peak whose frequency is below the threshold corresponds to a term in Wikipedia used in less than 0.01% of the cases. Since this term is highly rare, therefore, it can be ignored in the set of terms that form the middle-level. Indeed, it is recalled that the goal of the middle-level is to create a kind of set of elementary terms of the language.

Once this reduction has been made, the number of peaks decreases sharply, as shown in Table 6. The set of remaining peaks represents an approximation of the desired middle-level.

N. of final peaks	1.254
Peaks / synsets ratio	1,94%

Table 6: Number of identified peaks and ratio with synsets after frequency cut.

5.2 WSD: Test of Independence

In the previous section we have seen the Lesk algorithm for performing the word disambiguation task. Consider now a much less laborious process of disambiguation, in which each term is automatically assigned the first corresponding sense in WordNet, that is the most commonly used. This type of approach is certainly very efficient in terms of speed of execution and use of resources. On the other hand it is also true that it is much less refined compared to existing more advanced approaches, e.g., [5, 12]. We here present a comparison between the results obtained by following the two different approaches for the disambiguation of the senses (see Table 7).

#	With WSD	W/o WSD
Synsets	64.620	46.325
Noun synsets	53.191	40.387
Verb synsets	11.429	5.938
Leaves	72.271	70.908
Semantic trajectories	84.932	83.568
Peaks	14.190	15.369
Peaks after freq. cut	1254	1266
Overlapping	**91.70%**	**90.84%**

Table 7: Method's independence from WSD.

As one could imagine, the map of the sense frequencies created without disambiguating through the Lesk algorithm leads to having far fewer synsets than the previously analyzed version. Moreover, the number of detected peaks is considerably greater (in proportion to the number of synsets) in the second version. However, the size of the middle level remains almost equal, with a extremely high overlapping. This way, we proved that the proposed method is independent of the quality of the adopted WSD method.

5.3 Evaluation with Graded Readings

We can think of using the middle-level to make an assessment of the complexity of a text, perhaps for non-native speakers. To carry out this evaluation, three texts have been chosen with practically the same content but written in different levels of certified difficulties: the so-called *graduated readings*. Specifically, the three texts belong to the Elementary (A2), Intermediate (B1) and Upper Intermediate (B2) levels. These readings are available on the British Council website [2], which is a highly reliable and justifiable source for carrying out this type of test. The test consists in analyzing the three texts and verifying how many terms (or rather, senses) within these are part of the middle-level. Table 8 shows the results of this analysis

5.4 Frequent vs Middle-level Concepts

There are concepts that although very common do not fall into the middle-level. For example "*people*" turns out to be the thirty-ninth most frequent synset throughout Wikipedia, whereas in the middle-level it is not present. The 34th most frequent middle-level synset is in fact "*group*". Since "*group*" is hyponym of "*people*", our method always excludes the latter. In other words, "*people*" is a type of "*group*" (of the middle-level concept *person*) and it is the most used term to express a whole series of concepts, including "*people*", for which the latter is excluded. The same argument obviously applies to verbs. For example, consider "*stay*", in position 63 among the most frequent synsets. This is a hyponym of a sense of "*to be*". Therefore, even in this case, "*stay*" is excluded from the middle-level as its hyperonym is already part of it.

[2] http://britishcouncil.org

	Synset ID	*Lemma*	*Frequenza*
1	SID-02579744-V	be	42189707
2	SID-02182934-V	have, have got, hold	8369857
3	SID-02595485-V	constitute, represent, make up, comprise, be	3243027
4	SID-02536272-V	make, do	2753889
5	SID-02607558-V	Include	2471081
6	SID-01147708-V	use, utilize, utilise, apply, employ	2417515
7	SID-00147020-V	become, go, get	1936142
8	SID-14954729-N	old age, years, age, eld, geezerhood	1640694
9	SID-02725216-V	be	1606622
10	SID-07209466-N	time, clip	1581396
11	SID-01018451-V	name, call	1556292
12	SID-00341793-V	get down, begin, get, start out, start, set about, set out, commence	1553316
13	SID-01818343-V	travel, go, move, locomote	1449087
14	SID-02629830-V	be	1444912
15	SID-02591280-V	be	1396999
16	SID-06525881-N	movie, film, picture, moving picture, moving-picture show, motion picture, motion-picture show, picture show, pic, flick	1389362
17	SID-00999158-V	state, say, tell	1365122
18	SID-08095502-N	team, squad	1347747
19	SID-07968033-N	family, household, house, home, menage	1342090
20	SID-02246968-V	be	1295788
21	SID-00587430-V	know, cognize, cognise	1279608
22	SID-01089237-V	win	1239488
23	SID-15040391-N	season	1232568
24	SID-08379933-N	area, country	1217461
25	SID-01062243-V	play	1204267
26	SID-10153551-N	member	1189430
27	SID-02589075-V	be, live	1189164
28	SID-15004692-N	year, twelvemonth, yr	1185605
29	SID-02605487-V	bear	1156092
30	SID-02403508-V	establish, set up, found, launch	1138724
31	SID-06504272-N	album, record album	1111280
32	SID-13628130-N	part, portion, component part, component	1068174
33	SID-00451518-N	game	1000444
34	SID-00029714-N	group, grouping	987786

Figure 10: Top frequent middle-level peaks.

Level	Synsets (U)	Middle-level (U) ; %
A2	220 (114)	148 (73) ; **64%**
B1	279 (144)	186 (83) ; **58%**
B2	351 (190)	222 (102) ; **54%**

Table 8: Total and middle-level synsets within graded readings, with percentages. U = unique.

5.5 Further considerations

There is some aspect that has not been taken into consideration. Wikipedia contains not only conceptual and encyclopedic information, but also Named Entities (NEs). NEs are just entities with a

name, such as people, organizations, bands, music albums, movies, sports clubs, and so forth. Such entities are not really concepts and constitute a fairly large part in Wikipedia. This affects the construction of the middle-level: in fact, in the middle-level there are concepts such as *music albums* and *films* due to the high number of Wikipedia pages related to real music albums and films. That said, one could think in the future to filter the content of Wikipedia, perhaps before indexing, from the contents related to NE. Another point is related to the difference in learning language as a kid, or learning language as an adult. For instance, children tend to prefer concrete over abstract concepts. Thus, the use of a data source written by adults (Wikipedia) can reflect an adult-centric middle-level which could not entirely match with the language of children. For example, the method is shown to prefer '*publication*' over '*songbook*', even though for children, the latter would be preferred over the former, as the notion of a publication is not generally known to children).

6 RELATED WORK

The search of a set of basic concepts is not a new one, being approached under different perspectives. Many manually-drafted basic vocabulary lists have been proposed in the past, such as [2, 6, 11, 16] and others. However, the problem with most basic vocabulary lists is that they may be not representative, and polysemy is not managed [4]. Some studies have been instead focusing on how concepts may be defined by universal features [17]. Computational approaches and resources like WordNet [9] and EuroWordNet [15] proposed their own basic semantic classes. WordNet *super senses* are the roots of its taxonomy, representing top-level synsets only, thus not considering mid- and bottom-level concepts which can be undoubtedly part of the middle-level (e.g., "*school*"). EuroWordNet *base concepts* are synsets dominating several hyponyms. However, such selection reflects the specific structure of the lexical resource rather than the actual and crucial use of synsets.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a computational approach to identify what can be called *linguistic middle-level*, i.e., a subset of a language from which more complex communications develop in adulthood. Experiments on graded readings show promising results. This type of studies may have a significant impact on research in NL modeling and generation.

This research opens a large horizon of future perspectives and works. First, an evaluation and/or comparison among basic concept sets and vocabularies ([14]) can be considered in order to better highlight both shared/core concepts and differences. Then, the use of state-of-the-art WSD techniques such as [5, 12] can reveal better insights with respect to what found in Section 5.2. As already mentioned in Section 5.5, an improvement of the results can come from a finer selection of the Wikipedia content, for example by removing pages and sentences related to named entities. The employment of the simplified version of Wikipedia (https://simple.wikipedia.org/wiki/Main_Page) could additionally carry to interesting new findings. Still, as shown in Table 1, synsets can be considered middle-level concepts according to more than one semantic trajectory. This can be further studied in future work. Finally, one could consider more than one middle-level concept per semantic trajectory when the difference in terms of frequency is very small.

On the application side, graded and personalized language generation (e.g., [3, 8]) may be studied and developed on the top of this work. Text simplification ([13]) may exploit middle-level concepts to adapt texts at different levels of difficulty granularity. Finally, coarse-grained WSD tasks may be also considered for direct interaction and integration ([10]).

ACKNOWLEDGMENTS

The authors would like to thank Dr. Davide Maccagno for his work on the coding side. The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. Some of the future directions outlined in Section 7 come from their constructive feedback.

REFERENCES

[1] Roger Brown. 1958. How shall a thing be called? *Psychological review* 65, 1 (1958), 14.
[2] Edward W Dolch. 1936. A basic sight vocabulary. *The Elementary School Journal* 36, 6 (1936), 456–460.
[3] Ondřej Dušek and Filip Jurčíček. 2016. A context-aware natural language generator for dialogue systems. *arXiv preprint arXiv:1608.07076* (2016).
[4] Chan-Chia Hsu and Shu-Kai Hsieh. 2013. Back to the Basic: Exploring Base Concepts from the Wordnet Glosses. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 18, Number 2, June 2013-Special Issue on Chinese Lexical Resources: Theories and Applications.* http://www.aclweb.org/anthology/O13-3004
[5] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 897–907.
[6] David YW Lee. 2001. Defining core vocabulary and tracking its distribution across spoken and written genres: evidence of a gradience of variation from the British national corpus. *Journal of English Linguistics* 29, 3 (2001), 250–278.
[7] Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation.* ACM, 24–26.
[8] François Mairesse and Marilyn A Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics* 37, 3 (2011), 455–488.
[9] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.
[10] Roberto Navigli, Kenneth C Litkowski, and Orin Hargraves. 2007. Semeval-2007 task 07: Coarse-grained english all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations.* Association for Computational Linguistics, 30–35.
[11] Charles Kay Ogden. 1930. Basic English: A general introduction with rules and grammar. (1930).
[12] Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* 1156–1167.
[13] Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics* 165, 2 (2014), 259–298.
[14] Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016. Evaluating Lexical Simplification and Vocabulary Knowledge for Learners of French: Possibilities of Using the FLELex Resource.. In *LREC.*
[15] Piek Vossen. 1998. Introduction to eurowordnet. In *EuroWordNet: A multilingual database with lexical semantic networks.* Springer, 1–17.
[16] HE Wheeler and Emma A Howell. 1930. A first-grade vocabulary study. *The Elementary School Journal* 31, 1 (1930), 52–60.
[17] Anna Wierzbicka. 1996. *Semantics: Primes and universals: Primes and universals.* Oxford University Press, UK.