Strategies for Learning Lexemes Efficiently: A Graph-Based Approach

Jean-Marie Poulin, Alexandre Blondin Massé and Alexsandro Fonseca

Département d'informatique
Université du Québec à Montréal
Montréal, QC, Canada H3C 3P8
Email: poulin.jean_marie@courrier.uqam.ca

***Abstract*—Given a particular lexicon, what would be the best strategy to learn all of its lexemes? By using elementary graph theory, we propose a simple formal model that answers this question. We also study several learning strategies by comparing their efficiency on eight digital English dictionaries. It turns out that a simple strategy based purely on the degree of the vertices associated with the lexemes could improve significantly the learning process with respect to other psycholinguistical strategies.**

Keywords–Lexicons; Learning; Strategies; Graph theory.

I. INTRODUCTION

When learning a new language, the effort to develop a sufficient vocabulary basis plays an important role. Notwithstanding the fact that various cognitive skills are required, being able to associate a word with its meaning, its definition, is an essential part of the learning process. According to Schmitt [1], the "form-meaning link is the first and most essential lexical aspect which must be acquired". But as Gu and Johnson [2] mention, vocabulary acquisition is an intricate task. Joyce identifies and compares two such strategies, aimed at vocabulary improvement [3]: L_1 traduction, from the speaker's native language, and L_2 definition, in the language being learned. Traduction is in itself a very different problem, which we do not address here. As for the "L_2 definition" approach, it can be seen as the action of consulting a dictionary to acquaint oneself with the definition of a word in the new language, thus establishing this crucial "form-meaning link".

But what if this definition contains unknown words? Shall the reader examine in turn the definition of these unknown words in the lexicon? And then the unknown words in the definition of the unknown words, and so on? As discussed by Blondin Massé et al. in [4], this can lead to an infinite regression, the symbol grounding problem [5]. At some point in time, it is necessary to learn some words in ways other than dictionary perusal: either by sensorimotor experience, or through some other external contribution. In particular, it seems interesting to design a learning strategy to alleviate the burden of these potentially expensive forms of learning.

Dansereau characterizes a learning strategy as a sequence of "processes or steps that can facilitate the acquisition, storage and/or utilization of information" [6]. And in the more specific context of second language pedagogy, Bialystok defines it as "activities in which the learner may engage for the purpose of improving target language competence" [7]. One compelling alternative to the expensive direct learning approach is to identify a sequence of words, as small as possible, and ordered so as to minimize the overall learning effort: an "efficient learning strategy".

There is a large amount of related work aiming to identify small subsets of words from which one can learn all remaining words of a given language. It has been for a long time of great interest from psycholinguistic, pedagogical and computational points of view. For instance, in 1936, Ogden proposed a reduced list of 850 English words which would suffice to express virtually any complex words or thought [8]. In 1953, West [9] published the "General Service List" (GSL). Based on a corpus of 5 million words and containing about 2000 words, it is oriented toward the needs of students learning english as a second language. Despite having been criticized numerous times for its shortcomings and its incompleteness, it was considered until very recently as irreplaceable [10]. In the last few years, two principal contenders have been vying with one another to replace West's GSL. At about the same time in 2013, Brezina and Gablasova [11], and Browne [12] both presented what they call their New General Service List (NGSL), whose purpose is to restrict the attention to the most basic English words that should be understood first by non native speakers. A question remains open though: What is the optimal way to establish those word lists in an automated way?

All the word lists discussed above were built using large corpora. In a recent work, Nation [13] even describes a detailed corpus-based appoach to word list building. Our main contribution in this paper is to present a different, lexicon-based technique. To our knowledge, there has never been a fully computational, graph-based approach for identifying efficient learning strategies of a complete lexicon. Although in real life the process of learning words is clearly more intricate than the method we present here, our results suggest that an hybrid strategy, based both on cognitive observations and on formal tools such as graphs, could enhance significantly the way people improve their L_2 vocabulary.

The manuscript is divided as follows. In Section II, we introduce definitions and notation about lexicons, graphs and grounding sets. In Section III, we discuss different lexicon-based learning strategies. Section IV describes the data sets used in our analyses. Section V is devoted to the comparison of the efficiency of those different strategies. Finally, we briefly conclude in Section VI.

II. LEXICONS, GRAPHS AND GROUNDING SETS

We now propose formal definitions and notation about lexicons, when viewed as directed graphs. We believe that

 ISBN: 978-1-61208-609-5

this rather formal representation simplifies the discussion when comparing the efficiency of several lexicon learning strategies.

Roughly speaking, a *lexicon* can be defined as a set of lexical units, called *lexemes* enriched with definitions and arbitrary additional information [14]. For our purposes, we consider the following simplified representation of a lexicon.

Definition 1. A *lexicon* is a quadruple $X = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$, where

(*i*) $\mathcal{A}$ is a finite *alphabet*, whose elements are called *letters*.

(*ii*) $\mathcal{P}$ is a nonempty finite set whose elements are syntactic categories, called *parts-of-speech* (POS). In particular, it contains a special element denoted by STOP, which identifies lexemes whose semantic value is ignored.

(*iii*) $\mathcal{L}$ is a finite set of triples $\ell = (w, i, p)$ called *lexemes*, denoted by $\ell = w_p^i$, where $w \in A^*$ is a word form or simply word, $i \geq 1$ is an integer and $p \in \mathcal{P}$. If $p = \text{STOP}$, then ℓ is called a *stop lexeme*. We say that (w, i, p) is the *i-th sense of the pos-tagged word* (w, p). To make the numbering consistent, we also assume that if $(w, i, p) \in \mathcal{L}$ and $i > 1$, then $(w, i-1, p) \in \mathcal{L}$ as well. Whenever there exists $(w, i, p) \in \mathcal{L}$ with $i > 1$, we say that (w, p) and $\mathcal{L}$ are *polysemic*.

(*iv*) $\mathcal{D}$ is a map associating, with each lexeme $\ell \in L$, a finite sequence $D(\ell) = (d_1, d_2, \ldots, d_k)$, where $d_i \in A^*$ for $i = 1, 2, \ldots, k$, called the definition of ℓ.

If we replace the condition $d_i \in A^*$ by $d_i \in \mathcal{A}^* \times \mathcal{P}$ in (*iv*), then $D(\ell)$ is called a *POS-tagged definition* of ℓ and we say that X is a *POS-tagged lexicon*. It is also convenient to consider only the lemmatized, canonical form of words. If we replace in (*iv*) the condition by $d_i \in \mathcal{L}$, then $D(\ell)$ is called a *disambiguated definition* and we say that X is a *disambiguated lexicon*. Finally, if $D(\ell)$ is non empty whenever ℓ is a non-stop lexeme, then we say that X is *complete*.

Example 1. Let $X_1 = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$ be the lexicon such that

$$\begin{aligned} \mathcal{A} &= \{a, b, \ldots, z\}, \\ \mathcal{P} &= \{\text{N}, \text{V}, \text{A}, \text{R}, \text{S}\}, \end{aligned}$$

where N, V, A, R, S stand for *noun*, *verb*, *adjective*, *adverb*, STOP respectively, and $\mathcal{L}$ and $\mathcal{D}$ are both defined by Table I. Then X_1 is polysemic, lemmatized and disambiguated. Moreover, assuming that all words used in at least one definition are defined as well, X_1 is complete.

TABLE I. A DISAMBIGUATED LEXICON

Lexeme ℓ	$D(\ell)$
$fruit^1_{\text{N}}$	$(plant^1_{\text{N}}, part^1_{\text{N}}, that^1_{\text{S}}, have^1_{\text{V}}, seed^1_{\text{N}}, and^1_{\text{S}}, edible^1_{\text{A}}, flesh^1_{\text{N}})$
$fruit^2_{\text{N}}$	$(the^1_{\text{S}}, result^1_{\text{N}}, of^1_{\text{S}}, work^1_{\text{N}}, or^1_{\text{S}}, action^1_{\text{N}})$
$flesh^1_{\text{N}}$	$(the^1_{\text{S}}, edible^1_{\text{A}}, part^1_{\text{N}}, of^1_{\text{S}}, a^1_{\text{S}}, fruit^1_{\text{N}}, or^1_{\text{S}}, vegetable^1_{\text{N}})$
$flesh^2_{\text{N}}$	$(the^1_{\text{S}}, part^1_{\text{N}}, of^1_{\text{S}}, an^1_{\text{S}}, animal^1_{\text{N}}, use^1_{\text{V}}, as^1_{\text{S}}, food^1_{\text{N}})$
$seed^1_{\text{N}}$	$(the^1_{\text{S}}, small^1_{\text{A}}, part^1_{\text{N}}, of^1_{\text{S}}, a^1_{\text{S}}, plant^1_{\text{N}}, from^1_{\text{S}}, which^1_{\text{S}}, a^1_{\text{S}}, new^1_{\text{A}}, plant^1_{\text{N}}, can^1_{\text{S}}, develop^1_{\text{V}})$
$plant^1_{\text{N}}$	...
$etc.$	...

Lexicons are naturally converted to directed graphs. For a complete introduction to graph theory, the reader is referred to the classical book by Bondy and Murty [15], but for sake of self-consistency, we briefly recall some definitions and notation. The formal representation of lexicons is inspired from the definition of dictionaries found in [4].

A *directed graph* is an ordered pair $G = (V, A)$, where V is a finite set whose elements are called *vertices* and $A \subseteq V \times V$ is a finite set whose elements are called *arcs*. Directed graphs are useful for representing the relation "lexeme ℓ *defines* lexeme ℓ'": Given a disambiguated lexicon $X = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$, we define the *graph* $G(X)$ of X as the directed graph whose set of vertices is $V = \mathcal{L}$ and whose set of arcs A satisfies $(\ell, \ell') \in A$ if and only if $\ell \in D(\ell')$. In other words, the lexemes are the vertices, and there is an arrow from ℓ to ℓ' if and only if ℓ appears in the definition of ℓ'. Figure 1 depicts a subgraph of the graph $G(X_1)$ (see Example 1).

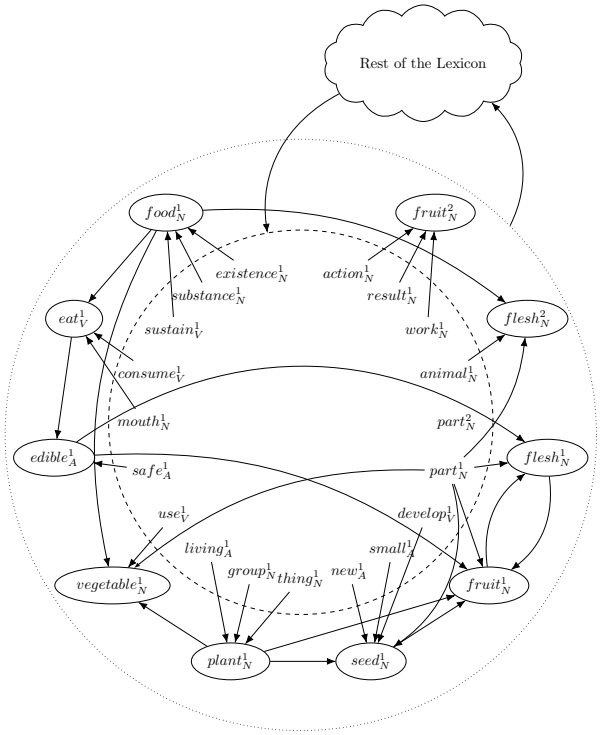

Figure 1. Graph of a polysemic, lemmatized, disambiguated, complete lexicon

Let $G = (V, A)$ be a directed graph. Given $u, v \in V$, we say that u is a *predecessor of* v if $(u, v) \in A$. The set of predecessors of v is denoted by $N^-(v)$ and the number of predecessors of v is called its *in-degree*, denoted by $\deg^-(v)$. Similarly, we say that v is a *successor of* u if $(u, v) \in A$, we denote by $N^+(u)$ the set of successors of u and we defined the *out-degree* by $\deg^+(u) = |N^+(u)|$. We define a map L on the subsets U of V by

$$L(U) = U \cup \{v \in V \mid N^-(v) \subseteq U\}.$$

From a linguistic point of view, $L(U)$ can be interpreted as the set of lexemes that can be *learned* from U, assuming that we can learn a new lexeme if and only if we already know it or if we know all lexemes appearing in its definition. In other words, L is a map associating with each set of lexemes U the set of lexemes $L(U)$ that can be learned directly from U. The

set U is called a *grounding set* of G if there exists a positive integer k such that $L^k(U) = V$, i.e., knowing U is sufficient to learn all remaining lexemes by definition alone in a finite number of steps.

We say that $p = (v_1, v_2, \ldots, v_k) \in V^k$ is a *path* of G if $(v_i, v_{i+1}) \in A$ for $i = 1, 2, \ldots, k-1$. If $v_1 = v_k$, then p is called a *circuit*. A *feedback vertex set of* G is a subset $U \subseteq V$ of vertices such that for every circuit c of G, the set $U \cap c$ is nonempty, i.e., U covers every circuit of G. It was proved in [4] that grounding sets are the same as *feedback vertex sets*, a well-known concept of graph theory. Unfortunately, the problem of computing feedback vertex sets is NP-hard for general graph, i.e., it is unlikely that one will ever find a general polynomial time algorithm solving the problem, as it has been shown by Karp [16]. Nevertheless, we were able to compute at least one minimum grounding set for most of our digital lexicons or find close approximations. According to Vincent-Lamarre et al,. it turns out that these minimum grounding sets present distinctive characteristics in comparison with other words: they are learned earlier, are more frequent and are slightly more concrete [17].

III. Learning Strategies

For sake of simplicity, we assume that there are essentially two complementary approaches for learning a new lexeme (see Harnad [5]): (*I*) *Directly*, i.e., by seeing, hearing, smelling, tasting, touching, or interacting in any other way with the object referenced by the lexeme; (*2*) *By definition*, i.e., by reading its definition or by having someone explain, describe, characterize the object with other lexemes.

We also make the following, quite strong assumptions in order to streamline the model: (*i*) Whenever the meaning of a lexeme is learned, it is learned permanently, i.e., it will never be forgotten; (*ii*) If we already know the meaning of all the lexemes occurring in the definition of some lexeme ℓ, then we can learn the meaning of ℓ simply by reading its definition; (*iii*) The effort of learning a lexeme *directly*, is more significant than the one of learning a lexeme *by definition*. (*iv*) Learning a lexicon efficiently amounts to learn its complete set of lexemes at a minimal cost. Taking into account the preceding assumptions, we define a learning strategy as follows.

Definition 2. Let $X = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$ a disambiguated lexicon. A *learning strategy of* X is any ordered sequence S whose elements are in $\mathcal{L}$. If X is a grounding set of X, then we say that S is *exhaustive*, otherwise we say that it is *partial*.

Simply stated, a learning strategy is a list of lexemes, ordered by decreasing priority. It is exhaustive if and only if it allows one to learn the complete lexicon. Intuitively, taking into account Assumption (*iii*), a learning strategy is efficient if its associated cost is as low as possible, which is realized exactly when the number of lexemes learned directly is minimal. Therefore, without loss of generality, we assume from now on that lexemes learned directly have cost 1, while lexemes learned by verbal instruction (definition) have cost 0.

The cost of a learning strategy can be handily computed. It is also easy to check if S is complete. In Algorithm 1, $\textsc{Cost}(S, X)$ returns an ordered pair (cost, X'), where *cost* is the cost of the learning strategy S for the lexicon X, and X' is the remaining part of the lexicon that could not be learned. Hence, S is complete if and only if X' is empty.

More precisely, Algorithm 1 proceeds as follows. First, it selects the next available lexeme in the strategy S and "learns it" with cost 1. Then, it learns all lexemes with no predecessor, i.e., lexemes whose definition contains only references to known lexemes, each at cost 0. This last step is repeated as long as there are available lexemes that can be learned at cost 0. Finally, it selects the next lexeme available in S and repeats the same process, until S is exhausted. The cost of using the strategy S is thus simply computed as the sum of the learning cost of all the lexemes in the lexicon.

Algorithm 1 Cost of a learning strategy

```
1:  function COST(S : strategy, X : lexicon) : (cost, lexicon)
2:      cost ← 0
3:      while S ≠ ∅ and X ≠ ∅ do
4:          ℓ ← S.POP()              ▷ We extract the next lexeme
5:          Remove ℓ from X          ▷ ℓ is learned at cost 1
6:          cost ← cost + 1
7:          while there exists ℓ ∈ X with deg⁻(ℓ) = 0 do
8:              Remove ℓ from X      ▷ ℓ is learned at cost 0
9:          end while
10:     end while
11:     return (cost, X)
12: end function
```

Any partial strategy S can easily be extended into a complete strategy S' by choosing a fallback strategy as soon as the list S of lexemes is exhausted. For instance, we could simply choose any random lexeme or choose a lexeme having a particular property, and repeat this process as long as there remain lexemes in the lexicon. Algorithm 2 presents such an extension by choosing, at each step, a lexeme whose number of occurrences in definitions is maximum.

Algorithm 2 Cost of a complete learning strategy

```
1:  function COMPLETECOST(S : strategy, X : lexicon) : cost
2:      (cost, X′) ← COST(S, X)
3:      while X′ ≠ ∅ do
4:          ℓ ← a lexeme of X′ such that deg⁺(ℓ) is maximal
5:          Remove ℓ from X          ▷ ℓ is learned at cost 1
6:          cost ← cost + 1
7:          while there exists ℓ ∈ X with deg⁻(ℓ) = 0 do
8:              Remove ℓ from X      ▷ ℓ is learned at cost 0
9:          end while
10:     end while
11:     return cost
12: end function
```

Both Algorithms 1 and 2 are efficient and easy to implement. More precisely, let n and m be respectively the number of vertices and arcs in the graph of the lexicon X. On one hand, Algorithm 1 has $\mathcal{O}(n+m)$ time complexity and $\mathcal{O}(n)$ space complexity, assuming that the removal of a single vertex is done in $\mathcal{O}(1)$, and by considering only *neighbors* of removed vertices when checking the condition in Line 7. On the other hand, Algorithm 2 runs in $\mathcal{O}(m \log n)$ time and $\mathcal{O}(n)$ space, with the same assumptions as for Algorithm 1, and by storing the candidate vertices in a priority queue. Indeed, in that case, Line 4 is done in $\mathcal{O}(\log n)$ time, and the time cost of all priority updates is $\mathcal{O}(m \log n)$, since each vertex v is updated in $\mathcal{O}(n)$ at most $\mathcal{O}(\deg(v))$ times.

 ISBN: 978-1-61208-609-5

It is obvious that some learning strategies are more efficient than other ones, given that the cost of a strategy depends strongly on the order in which the lexemes are organized.

IV. Data Sets

We now briefly describe the digital dictionaries and the psycholinguistic material for building the different learning strategies compared later in Section V.

Digital dictionaries. In our research, we construct and analyze 8 different digital lexicons, using dictionaries coming from 5 different sources. Two of the dictionaries, the *Longman's Dictionary of Contemporary English* (LDOCE) [18], and the *Cambridge International Dictionary of English* (CIDE) [19], are described by their authors as being built using a controlled vocabulary, using as few distinct words as possible in the definitions. The LDOCE is an advanced learner's dictionary, originally published in 1978. The CIDE is a dictionary originally developped in 1995 for advanced learners of English using the Cambridge Corpus. The 11th edition of the *Merriam-Webster's Collegiate Dictionary* (MWC) was published in 2003 [20]. With over 250 000 entries, it is by far the largest lexicon analyzed. *Wordsmyth* [21] is a linguistical educational project. It provides four different dictionaries: the *Wordsmyth Educational Dictionary-Thesaurus* (WEDT) was first developed in 1980, followed later by the *Wordsmyth Learner's Dictionary-Thesaurus* (WLDT), the *Wordsmyth Children's Dictionary-Thesaurus* (WCDT) and the *Wordsmyth Illustrated Learner's Dictionary* (WILD). The first two are targetted at adults, WEDT being for advanced learners and WLDT for beginners, while the last two are aimed at children. Finally, *WordNet* (WN) [22] is a well-known lexical network, whose purpose is not only to provide definitions of words, but also semantical relations between them, quasi-synonymy, antonymy and hypernymy being the most important. Table II presents statistics for the lexicons after preprocessing and removal of stop lexemes:

- The number of lexemes in each dictionary (#Lexemes);
- The number of POS-tagged lemmas (#Lemmas);
- The average number of senses by lemma (Polysemy);
- The number of lemmas actually used in definitions (#Lemmas used);
- The ratio of lemmas used over the total number of lemmas (#Usage ratio).

TABLE II. BASIC STATISTICS.

Lexicon	#Lexemes	#Lemmas	Polysemy	#Lemmas Used	Usage ratio
WILD	4 244	3 081	1.377	2 995	0.972
WLDT	6 036	3 433	1.758	2 212	0.644
WCDT	20 128	9 303	2.164	6 597	0.709
CIDE	47 092	18 694	2.519	8 773	0.469
LDOCE	69 204	22 511	3.074	10 074	0.448
WEDT	73 091	28 986	2.522	18 197	0.628
WN	132 547	57 243	2.316	29 600	0.517
MWC	249 137	68 181	3.654	33 533	0.492

The eight lexicons were then converted to disambiguated, graph-based lexicons, by using the Stanford's POS-tagger [23] and the "most frequent sense" heuristics, i.e., by choosing the most frequent sense each time it appears in some given definition. Graph structural statistics for the digital lexicons are shown in Table III:

- The number of nodes in the directed graph (#Nodes);
- The number of arcs in the directed graph (#Arcs);
- The number of strongly connected components (#SCCs);
- The size of the largest SCC (<SCC);
- The diameter of the largest strongly connected component (Diam.);
- The density of the graph (Density);
- The characteristic path length (CPL).

TABLE III. GRAPH STRUCTURAL STATISTICS.

Lexicon	#Nodes	#Arcs	#SCCs	<SCC	Diam.	Density	CPL
WILD	4 244	45 789	2 750	1 446	17	10.79	1.75
WLDT	6 036	28 623	5 088	858	25	4.74	1.10
WCDT	20 128	102 657	17 551	2 341	22	5.10	0.87
CIDE	47 092	334 888	45 306	1 702	16	7.11	0.21
LDOCE	69 204	415 052	67 224	1 770	16	6.00	0.16
WEDT	73 091	362 569	67 318	5 056	29	4.96	0.61
WN	132 547	694 067	124 589	7 079	30	5.24	0.50
MWC	249 137	1 155 085	239 478	8 842	29	4.64	0.31

Psycholinguistic variables. In order to build our learning strategies, we considered three different psycholinguistic variables: the *age of acquisition* (AOA), the *concreteness* (Conc) and the written and oral *frequency* (Freq). The *age of acquisition* variable indicates the age at which a word is first learned, on average. As references, we used two different sources. The first one is a database made available by Brysbaert. with words learned between the ages 1 and 21, with their surface forms and lemmas [24]. The second one comes from the *Child Language Data Exchange System* (CHILDES) project [25]. It contains transcripts of children's conversations with words learned between the ages 1 and 11. The *concreteness* variable indicates the level of materiality of a word, which varies from 1 (the less concrete/most abstract) to 5 (the most concrete). It was collected by asking participants to classify words into these categories [26]. For example: *banana*, *apple* and *baby* are level 5, *belief* is level 1.19 and *although* is level 1.07. Finally, the *frequency* variable corresponds to the rate of occurrence of words in a given corpus, normalized to one million. Brysbaert et al. used the SUBTLEX$_{US}$ corpus, as described in [27].

Derivated learning strategies. The number of learning strategies that one can design is huge. For a given dictionary containing n distinct senses, there are as many as $n!$, which is the number of permutations of the set $\{1, 2, \ldots, n\}$. We can distinguish two high-level categories:

1) *Graph-based* strategies are lists of lexemes built by exploiting the graph structure of a given lexicon. For instance, we could choose the next lexeme to learn by always picking the one whose out-degree is maximal (i.e., it appears in many definitions). It is worth mentioning that graph-based strategies are lexicon-dependent, i.e., they are adapted to the data.
2) *Psycholinguistic* strategies are obtained by choosing a lexeme according to its value with respect to a psycholinguistic variable. An example would be to pick first the lexemes that have been learned younger on average, up to the lexemes that have been learned later. Note that in that case, the strategies are lexicon-independent and are often incomplete, since psycholinguistic databases do not list all possible lexemes. Therefore, we need to complete the strategies by using a graph-based fallback strategy.

 ISBN: 978-1-61208-609-5

We focus our attention on 11 learning strategies that we now describe in more details. In the *minimum grouding set* strategy, the sequence of lexemes is built by picking minimum feedback vertex sets for each dictionary (see Section II). Although the problem is NP-complete in general, we were able to compute at least one optimal solution for 6 out of 8 dictionaries, and close approximations for the 2 remaining dictionaries. For the *dynamic degree* strategy, the list of lexemes is merely obtained by picking, at each step, the lexeme whose out-degree is maximal (since lexemes are "removed" at each step, the vertices degrees are indeed dynamic). From an algorithmic point of view, it corresponds to calling CompleteCost with S being the empty list. Concerning the *static degree* strategy, the list of lexemes is built beforehand in descending order of the vertices out-degree. This corresponds exactly to ordering the lexeme from the most frequently used in definitions to the less frequently used. The *Brysbaert/AOA, Brysbaert/Concreteness, Brysbaert/Frequency* and *Childes/AOA* are all "psycholinguistic strategies" built from [24]-[27]. A last strategy that we considered, called *NGSL/Frequency*, is obtained from the NGSL, designed precisely to help students learning English as a second language [12]. For the purpose of our analyses, this last list was enriched with 3 other lists provided by the same authors: the New Academic Word List (NAWL), the Business Service List (BSL), and the Technical Service List (TSL). In this last case, we also considered all the possible lemmas combinations as part of the strategy. For example, all the lemmas $something_{\text{NOUN}}$, $something_{\text{VERB}}$, $something_{\text{ADJ}}$, $something_{\text{ADV}}$ of the word *something* are included in the strategy. Finally, the strategies labelled *mixed* (*Grounding Set/Mixed*, *Dynamic Degree/Mixed* and *Static Degree/Mixed*) are described in Section V, as they are mostly used for comparing graph-based strategies with psycholinguistic strategies.

V. Results and Discussion

Table IV compares the efficiency of the learning strategies against each one of the 8 digital dictionaries. The following measures are shown:

- The total number of lexemes learned directly (Cost);
- The efficiency, which is the ratio of the total number of lexemes learned over the number of lexemes learned through direct learning (Effect). More precisely, if the efficiency of a strategy S for a lexicon X is e, then, on average, it costs 1 to learn e lexemes from X using strategy S.
- In a few cases, we also include the coverage (Cover), which is the percentage of lexemes that are learned before resorting to the fallback strategy (see Algorithm 2). Only those cases where the coverage is below 90% are shown.

The efficiency of the 8 English dictionaries is plotted in Figure 2. Note that, in both figures, the strategies are ordered in decreasing order of efficiency.

From Table IV and Figure 2, one can distinguish three groups, which are characterized by the speed with which they break "definition cycles". The first group consists only of the *minimum grounding set* strategy. Naturally, it is the most efficient, since it has been optimized in this regard. We include it mostly as a baseline for comparison with other strategies. The second group is composed of the *dynamic and static degree* strategies, plus the three mixed ones, are all graph-based, the

TABLE IV. LEARNING STRATEGIES EFFICIENCY. MGS: MINIMUM GROUNDING SET, DD: DYNAMIC DEGREE, SD: STATIC DEGREE, MGM: MINIMUM GROUNDING SET/MIXED, DDM: DYNAMIC DEGREE/MIXED, SDM: STATIC DEGREE/MIXED, NF: NGSL/FREQ., BF: BRYSBAERT/FREQ., CA: CHILDES/AOA, BA: BRYSBAERT/AOA, BC: BRYSBAERT/CONCRETENESS

Strat.	Meas.	CIDE	LDOCE	MWC	WN	WEDT	WCDT	WLDT	WILD
	Lexemes	47 092	69 204	249 137	132 547	73 091	20 128	6 036	4 244
MGS	Cost	349	484	1 544	1 251	1 365	570	231	340
	Effic	134.93	142.98	161.36	105.95	53.55	35.31	26.13	12.48
DD	Cost	684	843	3 095	2 566	2 389	897	394	574
	Effic	68.85	82.09	80.50	51.66	30.59	22.44	15.32	7.39
SD	Cost	687	838	3 081	2 558	2 386	899	394	577
	Effic	68.55	82.58	80.86	51.82	30.63	22.39	15.32	7.36
MGM	Cost	704	966	3 077	2 835	2 348	957	398	612
	Effic	66.85	71.64	80.96	46.75	31.13	21.03	15.17	6.93
DDM	Cost	768	963	3 466	3 002	2 574	987	448	645
	Effic	61.32	71.82	71.88	44.15	28.39	20.39	13.47	6.57
SDM	Cost	793	988	3 776	3 021	2 721	1 024	454	678
	Effic	59.32	70.00	65.98	43.87	26.86	19.65	13.30	6.25
NF	Cost	2 813	1 954	5 010	4 126	3 236	1 354	712	1 260
	Effic	16.74	35.42	49.73	32.12	22.59	14.87	8.48	3.37
	Cover.			**71.3%**	**73.4%**	**67.6%**	**82.9%**		
BF	Cost	6 751	2 170	8 217	7 204	6 555	1 999	960	1 193
	Effic	6.98	31.89	30.32	18.40	11.15	10.07	6.29	3.56
CA	Cost	4 971	5 010	7 729	7 284	5 586	3 409	1 585	2 016
	Effic	9.47	13.81	32.23	18.20	13.08	5.90	3.81	2.11
	Cover.			**82.9%**	**86.3%**	**84.3%**			
BA	Cost	7 105	4 851	10 119	10 340	8 278	2 950	1 284	1 430
	Effic	6.63	14.27	24.62	12.82	8.83	6.82	4.70	2.97
BC	Cost	8 900	11 669	16 580	17 037	12 792	6 042	2 373	2 477
	Effic	5.29	5.93	15.03	7.78	5.71	3.33	2.54	1.71

dynamic degree strategy being slightly better on average. In contrast, all strategies of the third group (*NGSL/Frequency*, *Brysbaert/Frequency*, *Childes/AOA*, *Brysbaert/AOA* and *Brysbaert/Concreteness*) can be qualified as "noisy". Indeed, they contain lexemes chosen and ordered based on some external criteria or psycholinguistic variables. Because of this specific ordering, many lexemes that could otherwise have been learned by definition are learned through direct learning and therefore increase the total cost. Hence, these strategies are not as efficient for definition cycle breaking.

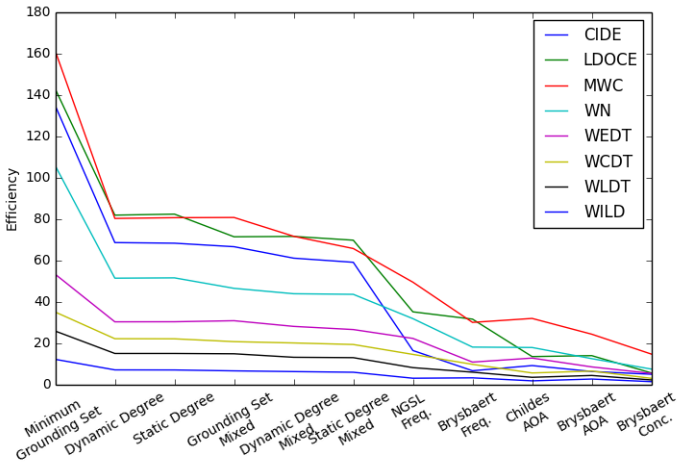

Figure 2. Dictionaries's efficiency

If we focus on the psycholinguistic strategies in the third group, it is worth mentioning that *NGSL/Frequency* turns out to be quite efficient, followed by *Brysbaert/Frequency*, both AOA-based strategies and, finally, *Brysbaert/Concreteness*. The fact that *Brysbaert/Concreteness* performs poorly is not very surprising. It seems quite natural to assume that one cannot learn a complete lexicon using only concrete lexemes: it is the combination of both concrete and abstract lexeme that

 ISBN: 978-1-61208-609-5

conveys a complete understanding of a lexicon. Although the AOA-based strategies do not perform well in comparison with other strategies, it still shows that knowing less than 10% of the lexemes is sufficient to learn all remaining ones by definition alone. A plausible explanation for this observation is that learning new lexemes cannot be done only by reading definitions. In other words, our model does not capture the whole learning process, which is not surprising. Another possibility is that the databases are not sufficiently complete and may contain errors which impact significantly the efficiency of the strategy.

Since the best strategies seem to be lexicon-dependent, one might wonder if an efficient lexicon-independent strategy could be designed or if the *NGSL/Frequency* strategy is optimal. This is not the case, as illustrated by three additional strategies called *Mixed MFVS*, *Mixed Dynamic* and *Mixed Static*. These lists have been obtained by merging strategies from the eight dictionaries, which implies that the *same list* is used for each dictionary. Although the performance decreases slightly when the lists are mixed in comparison with the nonmixed versions, they still all perform better than *NGSL/Frequency*.

VI. Concluding Remarks

In this paper, we introduced a new automated method for the construction of lexeme learning strategies. Instead of using a corpus or psycholinguistic variable, our approach is based on the internal structure of lexicons related to the domain of interest. We also described a formal model for representing lexicons and learning strategies, as well as related algorithms and metrics. These tools allowed us to quantitatively compare the overall performance of various strategies for learning complete lexicons.

In our experimentation with English language lexicons, we discovered that the most efficient strategies are those that quickly break definition circularity. In this regard, a simple strategy ordering the words according to the number of times they appear in other words definition turned out to be very efficient. Although we do not pretend that the value of word lists resides solely in their efficiency, we believe our approach is of interest, especially in situations where neither public word lists or large corpora in the domain of interest are available. In this case, the use of a digital lexicon or specialized dictionary would allow one to easily build a list of words or concepts pertinent to that domain, and above all, the order in which one should learn them.

We have many ideas to extend further our observations. For instance, we would like to study other lexemes learning strategies, either based on the graph structure of the lexicon or from psycholinguistic variables. It would also be interesting to apply our model to lexicons specialized for a particular field, such as mathematics, medical care, computer sciences, etc. Finally, it seems reasonable to expect that our observations are language independent. However, this is harder to verify since the databases available in languages other than English are often less complete.

References

[1] N. Schmitt, "Instructed second language vocabulary learning," *Language teaching research*, vol. 12, no. 3, pp. 329–363, 2008.

[2] Y. Gu and R. K. Johnson, "Vocabulary learning strategies and language learning outcomes," *Language Learning*, vol. 46, no. 4, pp. 643–679, 1996.

[3] P. Joyce, "L2 vocabulary learning and testing: The use of l1 translation versus l2 definition," *The Language Learning Journal*, pp. 1–12, 2015.

[4] A. Blondin Massé, G. Chicoisne, Y. Gargouri, S. Harnad, O. Picard, and O. Marcotte, "How is meaning grounded in dictionary definitions?" in *TextGraphs-3*. Morristown, NJ, USA: Association for Computational Linguistics, 2008, pp. 17–24.

[5] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.

[6] D. F. Dansereau, "Learning strategy research," *Thinking and learning skills*, vol. 1, pp. 209–239, 1985.

[7] E. Bialystok, "Some factors in the selection and implementation of communication strategies," *Strategies in interlanguage communication*, pp. 100–118, 1983.

[8] C. K. Ogden, *Basic English: A general introduction with rules and grammar*. K. Paul, Trench, Trubner, 1944, vol. 29.

[9] M. West and M. P. West, *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Addison-Wesley Longman Ltd, 1953.

[10] A. Coxhead, "A new academic word list," *TESOL Quarterly*, vol. 34, no. 2, pp. 213–238, July 2000.

[11] V. Brezina and D. Gablasova, "Is there a core general vocabulary? introducing the new general service list," *Applied Linguistics*, vol. 36, no. 1, pp. 1–22, 2013.

[12] C. Browne, "A new general service list: The better mousetrap we've been looking for," *Vocabulary Learning and Instruction*, vol. 3, no. 2, pp. 1–10, 2014.

[13] I. S. Nation, *Making and using word lists for language learning and testing*. John Benjamins Publishing Company, 2016.

[14] M. Aronoff and J. Rees-Miller, *The handbook of linguistics*, 1st ed. John Wiley & Sons, 2002.

[15] J. A. Bondy and U. S. R. Murty, *Graph Theory with Applications*. New York: Elsevier, 1976.

[16] R. M. Karp, *Reducibility among Combinatorial Problems*. Boston, MA: Springer US, 1972, pp. 85–103.

[17] P. Vincent-Lamarre, A. Blondin Massé, M. Lopes, M. Lord, O. Marcotte, and S. Harnad, "The latent structure of dictionaries," *TopiCS: Discovering principles of cognition by mining large, real-world data sets*, 2016.

[18] P. Procter, *Longman Dictionary of Contemporary English (LDOCE)*. Essex, UK: Longman Group Ltd., 1978.

[19] ——, *Cambridge International Dictionary of English (CIDE)*. Cambridge University Press, 1995.

[20] *Merriam-Webster's Collegiate Dictionary*, 11th ed., 2003.

[21] Wordsmyth, [Retrieved: January 2018]. [Online]. Available: https://www.wordsmyth.net

[22] C. Fellbaum, Ed., *WordNet An Electronic Lexical Database*. Cambridge, MA ; London: The MIT Press, 1998.

[23] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *NAACL 2003*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 173–180.

[24] V. Kuperman, H. Stadthagen-Gonzalez, and M. Brysbaert, "Age-of-acquisition ratings for 30,000 english words," *Behavior Research Methods*, vol. 44, no. 4, pp. 978–990, 2012.

[25] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates, third edition, 2000.

[26] M. Brysbaert, A. B. Warriner, and V. Kuperman, "Concreteness ratings for 40 thousand generally known english word lemmas," *Behavior research methods*, vol. 46, no. 3, pp. 904–911, 2014.

[27] M. Brysbaert and B. New, "Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english," *Behavior research methods*, vol. 41, no. 4, pp. 977–990, 2009.

 ISBN: 978-1-61208-609-5