The Dictionary Game: Toward a Characterization of Lexical Primitives Using Graph Theory and Relational Concept Analysis

Mickaël Wajnberg, Jean-Marie Poulin, Alexandre Blondin Massé and Petko Valtchev

Département d'informatique, Université du Québec à Montréal, Montreal, Quebec, Canada H3C 3P8
Emails: wajnberg.mickael@courrier.uqam.ca poulinjm@gmail.com
blondin_masse.alexandre@uqam.ca valtchev.petko@uqam.ca

***Abstract*—In language theory and cognition, the search for a minimal set of language primitives from which every other concept could be defined is an ever-recurring topic. In order to help identify such primitives, a serious game was designed, where the player has to produce a meaningful lexicon as small as possible, starting by defining a single word and, recursively, all those appearing in any definition. Using simple graph theory and relational concept analysis (RCA), we extracted association rules from feature tables while putting in common the newly discovered abstractions into the overall knowledge data discovery process. The utility of the mined rules has been validated by the success in linking the dictionaries structural attributes to the psycholinguistics characteristics of the words they contain.**

Keywords–Lexicon; Dictionary; Relational concept analysis; Cicularity; Association rule; Serious game.

I. INTRODUCTION

If someone is trying to learn a new language using only a dictionary, he must first identify a set of words in the foreign language's dictionary that he can relate to words in his mother tongue. one must also ensure that this set of words covers the lexical primitives (assuming that they exist) of the foreign language, *i.e.*, a set of "indecomposable" words sufficiently large to span all the other words of the language. The process of acquiring the first words of the foreign dictionary is addressed by the so-called *symbol grounding problem*, which was formalized in 1990 [1]. The task is all the more difficult when the alphabets do not match, like Mandarin or Arabic. It is nevertheless achievable, and even characteristic of the work of palaeographers aiming to understand extinct languages.

To identify such lexical primitives of a language, some authors put forward the concept of *minimal grounding sets* (MGS) of dictionaries, *i.e.*, minimum size sets of dictionary words from which one can define all the other words in a dictionary [2][3]. To properly characterize these MGS, it is necessary to understand both their structural aspect and their description from a psycholinguistic point of view [3].

By means of *association rules*, we analyze in this article a body of small dictionaries produced by human players as an artefact of a serious game called "Dictionary Game". These association rules are derived using a mathematical procedure known as *Relational Concept Analysis* (RCA) [4], an extension of *Formal Concept Analysis* (FCA) [5]. First, in Section II, we detail the context of the study. Section III introduces the formalization of a dictionary. Next, in Section IV, we describe in more detail "The Dictionary Game", which supplies our datasets. Finally, Section V is devoted to the description of the characteristics of the data set and the experiment carried out.

II. SYMBOL GROUNDING AND LEXICAL PRIMITIVES

In language theory, as well as in cognitive science, the search for lexical primitives has been a very active subject for several decades [6][7]. These primitives form a set of lexical units, such that any word in a language can be defined from them. In theory, we can integrate in an iterative way this collection to eventually define all the words of a dictionary. To be really helpful in this purpose, a valuable set of lexical primitives should both contain as few words as possible and be as expressive as possible.

One of the famous first attempt to identify a minimum set of lexical primitives was made in 1930 by Ogden [8]. Although he did not succeed in constructing a universal language, one can still retrieve his word list [9]. Emphasizing the timeless aspect of this line of research, the graph structure of Ogden's Basic English Word List was even recently studied [10].

In 1972, Wierzbicka introduced a group of 14 semantic primitives, which she considers to be universal *Semantic Primitives* [11]. Pursuing this line of research for two decades, she extended her list of words to more than 50 semantic primitives and has shown that they can be translated into a large number of languages [12]. Not long ago, Browne et *al.* have constructed several "general" lists of words, such as the *New General Service List* [13][14]. These word lists, carefully chosen to meet the requirements of lexical primitives, have been used primarily for teaching English, but have also been used in other contexts [15]. More recently, Goddard proposed a *Minimal English* based on Wierzbicka's theory [16]. It aims to manually build a set of basic English words which allow to describe a large number of more complex words, which are translatable into many different languages [16]. Its main goal is to provide a basic language as an effective entry point for learning English as a secondary language.

Although useful in practice, all these approaches start by proposing a set of primitives and try, from these, to construct as many concepts as possible. However, the authors of these methods emphasize that they should not be considered as complete and definitive [17]. Even the notion of "indecomposability" is not obvious: Wierzbicka's list of 65 primitives includes the concepts NOT, GOOD and BAD; however, one could argue that among the concepts GOOD and BAD, at least one of them could be removed, since each can be defined by combining the other with the negation concept NOT, *i.e.*, GOOD = NOT BAD and BAD = NOT GOOD.

Still in response to the problem of symbol grounding, a complementary approach was developed by Blondin Massé et al. [2]. As an alternative to a fixed set of word primitives, the authors propose to calculate the MGS for dictionaries, *i.e.*, identify in a dictionary a minimum set of words which allow to define in an iterative manner all the other words. Although the number of MGS for a dictionary can be exponential with respect to its number of words, all MGS seem to share common psycholinguistic characteristics [3]. Indeed, it has been shown that these words tend to be used more frequently, learned younger and are more concrete [3]. For these psycholinguistic variables (age of acquisition, frequency, concreteness), one can also note a difference, depending on the part of the dictionary in which a word is found.

When taking a closer look at several MGS, other interesting features have been observed. Indeed, the authors of [3] focused on the largest strongly component connected (SCC), called the *core* of a dictionary, and compared it to the smallest remaining SCC, called the *satellites*. They found out that, when partitionning the MGS into two parts, one in the core and the other in the satellites, the words in the core are more frequent, more abstract and learned earlier, unlike those found in the satellites [3]. Thus, the core seems to mirror some abstraction occurring in the mental lexicon.

III. Preliminaries

We now recall the terminology about lexicons and graphs. Formalism on lexicons is adaptated from [18].

Definition 1. [18] A *complete disambiguated lexicon* is a quadruple $X = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$ where

- $\mathcal{A}$ is a finite *alphabet*, whose elements are *letters*;
- $\mathcal{P}$ is a finite set whose elements are *part-of-speech (POS)*;
- $\mathcal{L}$ is a finite set of triples of the form $\ell = (w, i, p)$, called *lexemes*, denoted by $\ell = w_p^i$, where $w \in A^*$ is a word, $i \geq 1$ is an integer and $p \in P$. The triple (w, i, p) is called the i*-th sense* of the *POS-tagged* ordered pair (w, p)
- $\mathcal{D}$ is a partial application associating with a lexeme $\ell \in \mathcal{L}$ a finite non empty sequence $D(\ell) = (\ell_1, \ell_2, \ldots, \ell_k)$, where for each i, $d_i \in \mathcal{L}$. Such a sequence is called the *definition* of ℓ.

The quadruple must satisfy the following constraints:

stop lexeme The set $\mathcal{P}$ contains a special element S, identifying the *stop lexemes*;

completeness For each triple $(w, i, p) \in \mathcal{L}$, if $p \neq S$, then $\mathcal{D}(w, i, p)$ is well-defined;

consistent numbering of lexemes If $(w, i, p) \in \mathcal{L}$ and $i > 1$, then $(w, i - 1, p) \in \mathcal{L}$.

If for each triple $(w, i, p) \in \mathcal{L}$, we have $i = 1$, then the lexicon $X = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$ is called *monosemic*. In that case, we write w_p instead of w_p^i.

TABLE I. A complete disambiguated lexicon.

ℓ	$D(\ell)$
BIG$_A$	(NOT$_S$, SMALL$_A$)
HUGE$_A$	(VERY$_S$, BIG$_A$)
SMALL$_A$	(NOT$_S$, BIG$_A$)

Roughly speaking, a complete disambiguated lexicon is a list of lexemes that are all defined, except the stop lexemes, and such that each definition is disambiguated.

Example 1. Let $X = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$, where $\mathcal{A} = \{a, b, \ldots, z\}$, $\mathcal{P} = \{A, S\}$ (N = name, A = adjective, S = stop) and $\mathcal{L}$, $\mathcal{D}$ are defined in Table I. Each lexeme used in a definition is itself defined, except the stop lexemes NOT$_S$ and VERY$_S$. Hence, the lexicon X is complete and disambiguated.

In practice, words tagged with S are words playing mostly a syntactic role and whose semantic value is poor (such as *no*, *the*, *a*). However, any word can be placed in that category whenever its sense is not pertinent for a given study. From now on, we assume that $\mathcal{P} = \{N, V, A, R, S\}$, denoting respectively the POS *name*, *verb*, *adjective*, *adverb* and *stop*.

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*. The *density* of G, denoted by $\mathrm{density(G)}$, is the ratio of the number of arcs belonging to the graph over the number of possible arcs, *i.e.*, $\mathrm{density(G)} = |A|/|V|^2$.

Let $G = (V, A)$ be a graph, $u, v \in V$ and k be a positive integer. We say that $p = (v_1, v_2, \ldots, v_k)$ is a *(directed)* uv*-path* of G if $u = v_1$, $v = v_k$ and $(v_i, v_{i+1}) \in E$ for $i = 1, 2, \ldots, k - 1$. In particular, if $u = v$, the path p is called a *circuit* of G. Let $u, v \in V$. We write $u \rightarrow_G v$ whenever there exists a uv-path in G, or simply $u \rightarrow v$ if the graph G is clear from the context. Also, we write $u \leftrightarrow v$ if and only if $u \rightarrow v$ and $v \rightarrow u$. It is easy to verify that $\leftrightarrow$ is an equivalence relation. Hence, an equivalence class of the relation $\leftrightarrow$ is called *strongly connected component (SCC)* of $G = (V, A)$. In other words, two vertices belong to the same SCC if there exist directed paths connecting the first one to the second one and vice-versa. SCC can be computed in linear time by different algorithms, such as Tarjan's [19].

When computing statistics about directed graphs, it is often convenient to consider their undirected version. Given two vertices u, v of a directed graph $G = (V, A)$ and a positive integer k, we say that $p = (v_1, v_2, \ldots, v_k)$ is a uv*-chain* of G if $u = v_1$, $v = v_k$ and, for each $i = 1, 2, \ldots, k - 1$, we have $(v_i, v_{i+1}) \in E$ or $(v_{i+1}, v_i) \in E$. The *length* of a uv-chain, denoted by $|p|$, is the number $k - 1$, *i.e.*, the number of arcs traveled by p. The *distance between* u *and* v, denoted by $\mathrm{dist}(u, v)$, is the length of a shortest chain between u and v, *i.e.*,

$$\mathrm{dist}(u, v) = \min\{|p| : p \text{ is a } uv\text{-chain}\}.$$

From these definitions, we can derive structural statistics for a given graph G. For instance, the *diameter of* G, denoted by $\mathrm{diam}(G)$, is the maximal distance between two vertices of G:

$$\mathrm{diam}(G) = \max\{\mathrm{dist}(u, v) : u, v \in V\}.$$

Finally, the *characteristic path length (CPL)* of G is the average length of a shortest path between two vertices. It is denoted by $\mathrm{CPL}(G)$ and defined by

$$\mathrm{CPL}(G) = \sum_{u,v \in V} \frac{\mathrm{dist}(u,v)}{|V|(|V|-1)}.$$

If $X = (\mathcal{A}, \mathcal{P}, \mathcal{L}, \mathcal{D})$ is a complete disambiguated lexicon, then the *graph* of X is the directed graph $\mathrm{Graph}(X) = (\mathcal{L}, A)$, such that $(\ell_1, \ell_2) \in A$ if and only if the lexeme ℓ_1 appears in the définition $\mathcal{D}(\ell_2)$ of lexeme ℓ_2.

IV. The Dictionary Game

The "Dictionary Game" is a web-based, crowdsourced game, whose purpose is to create small but complete micro dictionaries of tractable size [3][20].

The reader can take a look at the game's web site for a more complete description of the game and even get down to build a new dictionary of one's own [21]. At the outset of the game, the player has to pick a "seed word" and provide a definition for it. After that, the words used in this first definition must in turn be defined. The game continues in the same manner, new definitions being created, new words being added and defined using existing or new words. The goal of the game is to "complete" the dictionary so that all the words used in definitions are themselves defined. Thus, the dictionary obtained at the end verifies all criteria of a complete lexicon according to Definition 1. To improve the expressiveness of the resulting dictionary, the player must also ensure that all the definitions provided contain at least three non stop lexemes. An error message is displayed if this constraint is not satisfied, inviting the player to improve the definition. Example 2 shows the written representation of the first words and definitions of a dictionary built using the seed word *horse*.

Example 2. Using the seed word *horse*:

- *horse*: animal on which one human rides
- *animal*: organism that belongs to the living kingdom
- *human*: animal species that possess reason
- etc.

Also, a graphical representation of the graph associated with a dictionary produced from the seed word *clock* is depicted in Figure 1.

Some additional data representation aspects must be taken into account to prepare the output of the "Dictionary Game" for further analysis.

1) Seed Words: As mentioned just before, game dictionaries are built using seed words. In the current version of the game, one can choose between 4 different seed words. We can see in the first columns of Table V these seed words along with the number of dictionaries built for each of them.

2) Graph Characteristics: The dictionaries produced by the players are converted to directed graphs, using the natural transformation described in Section III. To get an overview of the underlying structure, we computed several classic measures on the resulting graphs, summarized in Table II.

3) Words Psycholinguistic Properties: In order to portray the words used to build the game lexicons, we used external norms to tag them according to their psycholinguistic properties. Table III shows a few sample words along with their associated psycholinguistic properties:

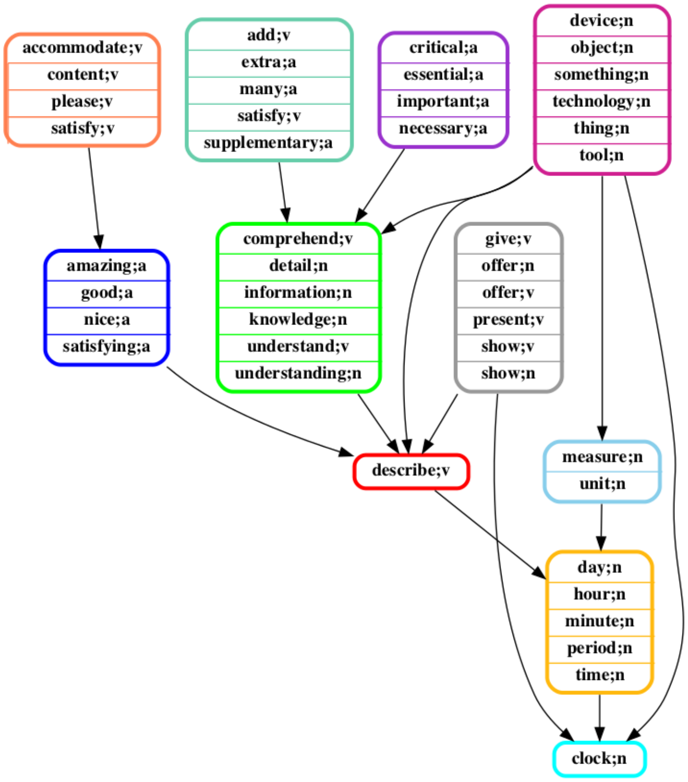

Figure 1. A dictionary produced by a player in the game, represented as a so-called "condensed" graph. Each SCC has been merged into a single meta-vertex containing equivalent words with respect to the relation ↔.

TABLE II. Statistical properties of graph characteristics for all game dictionaries: *numV*: number of vertices, *numE*: number of edges, *nSCC*: number of Strongly Connected Components, *CPL*: characteristic path length, *dens*: density, *diam*: diameter.

	numV	*numE*	*nSCC*	*CPL*	*dens*	*diam*
Mean	124.0	436.1	7.0	4.8	0.038	13.1
Std	73.9	272.7	11.2	1.5	0.022	4.7
Min	35	120	1	0.87	0.008	3
Max	433	1558	88	11.4	0.108	35

Age of Acquisition: The variables AOAB and AOAC both represent estimations of the age at which a word is supposed to have been learned. They were sourced from psycholinguistic norms, [22] and [23], respectively.

Concreteness: The variable *Conc* is an evaluation, on a 1 to 5 scale, of whether a word is abstract or concrete [24].

Frequency: *FreqP* – [25] is a measure of the relative occurrence rate of words in the SUBTLEX$_{US}$ corpus, while

TABLE III. A sample of words and their psycholinguistic properties. A missing value is written as a dash.

	AOAB	AOAC	*Conc*	*FreqP*	*FreqL*
abandon	8.32	—	2.54	8.10	1
abide	9.50	4.00	1.68	2.71	1
ability	8.84	—	1.81	19.22	38
able	7.79	4.77	2.38	159.90	39
absence	7.70	—	2.31	6.31	5
absent	6.50	8.28	2.70	2.57	1
...					

FreqL is a measure of the words' frequency in the corpus formed by collecting all the words from the game dictionaries definitions.

V. Experiments

We now describe the experiment we conducted, the resulting observations, and end with a short discussion.

A. Objectives and Method

The goal of this study is to understand the underlying structure of the game produced dictionaries. Specifically, since the game asks the players to construct dictionaries with a minimal number of words, we focused on the "winning" strategies. To provide such insights, we aim to extract co-occurrences between psycholinguistic and structural features in the dictionaries. To present these co-occurrences we use a dedicated formalism, the *association rules* [26]. Such rules are pairs $Y \rightarrow Z$, where Y and Z are sets of features respectively called *antecedent* and *consequent*, and state that any object presenting all the Y features, has also all the Z features. Associations are typically assessed by metrics, such as the *rules support* (proportion of objects incident to $Y \cup Z$) and *confidence* (proportion of objects with Z among those with Y). For clarity, we also provide the *antecedent support*, the proportion of objects presenting all the antecedent features. For instance, Rule #4 in the Table VII states that 35% objects (dictionaries) are $numV(lo)$ (low number of vertices), 29% objects are $numV(lo)$ and $dens(hi)$ (high density), and finally, 84% objects that are $numV(lo)$ are also $dens(hi)$.

To limit redundancy and maximize informativity, we focus on associations of a special form, called *concise association*. Such association are written $Y \rightarrow Z - Y$, where $Y \subseteq Z$ and there is no sets of features U, V such as $U \subseteq Y$, $Z \subseteq V$, where the rule $U \rightarrow V$ has the same support or confidence than $Y \rightarrow Z - Y$ [27]. To produce these specific associations rules, we use *RCA* [4]. FCA is a method that reveals the concise association rules of objects × attributes datasets (called *formal context*), such as the table *dict* in Figure 2, by expliciting, in a lattice, the hidden conceptual structure [5]. RCA extends FCA to the case where relations exists between objects, such as described in Example 3. The input of RCA is called a *Relational Context Family* (RCF), *i.e.*, a pair composed of a set of contexts and a set of binary relations between these concepts.

Example 3. Consider Figure 2. The *dict* table depicts a set of the dictionaries 1 to 5 with 6 features : three levels of vertices number $numV(lo)$, $numV(med)$ and $numV(hi)$ along with three seed words $horse$, $clock$ and $person$. Crosses indicate that the object has the given attribute, so dictionary 1 has many vertices and has been started with the seed word $clock$. Such a table, composed with a set of *objects* (the dictionaries), a set of *attributes* (the features) and an *incidence relation* (the set of couples represented by the crosses) is called a *formal context*. The *wd* table presents the *formal context* of words A to D with the attributes "young Brisbaert age of acquisition", "high concretude" and "lowest P-frequency". These two contexts can both be analyzed separately through FCA. RCA enrich each context with the use of *relations*, such as the one represented by the *ct* table (*ct* stands for "contains"), linking the dictionaries to the words by specifying which word exists in which dictionary.

dict	*numV*(hi)	*numV*(med)	*numV*(lo)	horse	clock	person
1	×				×	
2			×		×	
3			×			×
4	×			×		
5		×			×	

ct	*A*	*B*	*C*	*D*
1	×		×	×
2		×	×	
3				×
4	×	×	×	×
5	×		×	

wd	AOAB(young)	Conc(hi)	FrP(q_1)
A	×		
B	×	×	
C		×	
D	×		×

Figure 2. A sample RCF drawn from the dictionary game dataset.

TABLE IV. Extended problem context.

$dict{+}$	1	2	3	4	5
$\forall\exists ct$: ($\top$)	×	×	×	×	×
$\exists_{60\%} ct$: (AOA(young))	×		×	×	
$\forall\exists ct$: (AOA(young), *FrP*(q_1))			×		
$\forall\exists ct$: (Conc(hi))		×			
$\forall\exists ct$: (AOA(young), Conc(hi))					

On such data, RCA iteratively performs multiple FCA tasks, one per context in the RCF. In doing that, relational links between objects are translated into special type of attributes, called *relational*, by a dedicated *propositionalization* mechanism [28]. It applies a variety of *scaling operators* mimicking role restrictions from description logics thus yielding attributes of the shape $q\ r : (Y)$ where q is the operator (*e.g.*, $\exists$, $\forall$, $\forall\exists$), r is a relation name (*e.g.*, *ct*) and Y is a set of attributes from the range context of r (*wd* for r =*ct* in our RCF). To avoid circularity and reduce redundancy, only maximal sets of attributes computed in anterior RCA iteration are considered, as suggested in [29]. An attribute $q\ r\ :\ (Y)$ refines the descriptions of the objects from the domain context of r (*dict* for *ct*), *i.e.*, it becomes an additional column in the ×-table, whereby its incidence to an object o is function of object's image, $r(o) = \{\bar{o} | (o, \bar{o}) \in r\}$, and the objects having the attributes Y in the range context. The exact function is defined by the operator q, *e.g.*, $\forall\exists$ tests $r(o) \neq \emptyset$ and if every objects of $r(o)$ has Y while $\cap_{\geq 60\%}$ checks if at least 60% objects of $r(o)$ has Y. Table IV presents some of the attributes generated by scaling upon *ct* with operators $\forall\exists$ and $\cap_{60\%}$.

As a result of the scaling, dictionary descriptions is refined with respect to the properties of the words they comprise, *e.g.*, 2 and 3 are both small-sized, yet 2 contains only highly concrete words which is not true for 3. When RCA terminates, a last FCA task is launched to generate the association rules from the final and extended contexts.

B. Model

As presented, RCA allows association rules extraction on a ×-table dataset. Therefore, dictionaries and words attributes need to be discretized into categorical attributes to enable the use of this method. For example, we can observe in Table V the

TABLE V. EXAMPLE OF DISCRETIZATION FOR THE *numV*GRAPH CHARACTERISTIC: *Seed*: SEED WORD USED TO BUILD THE DICTIONNARY, *NbDicts*: NUMBER OF DICTIONARIES BUILT USING THE SEED WORD, *Lo/Med/Hi* UPPER/LOWER LIMITS USED TO ESTABLISH THE CATEGORY

Seed	NbDicts	Lo	Med	Hi
clock	47	[39, 71]	[72, 111]	[112, 433]
horse	24	[40, 107]	[108, 136]	[137, 407]
person	13	[40, 108]	[109, 158]	[159, 243]
thing	10	[35, 103]	[104, 162]	[163, 288]

TABLE VI. DISCRETIZATION OF PSYCHOLINGUISTIC VARIABLES

Words		
Property	Range	Category
AOAB	[2.3, 6.6]	young
	[6.7, 9.2]	middle
	[9.3, 16.2]	older
AOAC	[1.3, 3.74]	young
	[3.75, 4.70]	middle
	[4.71, 11.0]	older
CONC	[1.1, 2.3]	lo
	[2.4, 3.6]	med
	[3.7, 5.0]	hi
FREQP	[0.02, 3.56]	lo
	[3.57, 25.52]	med
	[25.53, 6161.41]	hi
FREQL	[0, 1]	Q1
	[2, 3]	Q2
	[4, 8]	Q3
	[9, 87]	Q4

minimum and maximum values for discrete categories for the *numV* property which equates to the number of vertices. For the *clock* seed word, the dictionaries whose *numV* is less than 72 are assigned to category *lo*, to category *med* if it is 72 or more but less than 111, and to category *hi* if it is 112 or more. We proceeded in a similar manner to subdivide the words into categories according to the value of their psycholinguistic properties, as shown in Table VI.

Using this discretization, we designed an RCF such as the one presented in Figure 3. The *words* formal context describes the words present in at least one of the dictionary with the discretized attributes presented in Table VI. The *dict* formal context describes the dictionaries using the seed words and the discretization, as shown in Table V, of every structural variable presented in Table II. Along with these two contexts, we use the relations *contains* (ct) that specify which words a dictionary contains and the inverse relation *in* that indicates in which dictionary a word is. An excerpt is presented in Figure 2.

To highlight special word classes, the relation *in* is scaled with the propositionnalization operator $\forall\exists$. A word having an attribute $\forall\exists in : (Y)$ can be interpreted as being used exclusively in dictionaries presenting all the features of Y. On the other side, the relation the relation *in* is scaled with the propositionnalization operators $\cap_{\geq p\%}$ for $p \in \{40, 50, 60, 70, 80, 90, 100\}$. A dictionary having an attribute $\cap_{\geq p\%}$ can be interpreted as being composed of at least $p\%$ words presenting all the features of Y. Note that $\cap_{\geq 100\%} = \forall\exists$. This choice is done because a granularity of 0.1 seems sufficient and below 40% an attribute does not present relevant information (characterizing a dictionary that has at least one word learned at young age does not bring substantial information).

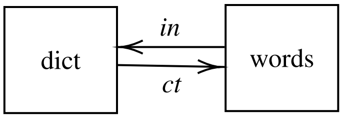

Figure 3. RCF schema used in our experiment

TABLE VII. SOME ASSOCIATION RULES PRODUCED BY THE RCF OF FIGURE 3

#	Antecedent	Consequent	Antecedent Support	Rule Support	Confidence
1	$\cap_{\geq 60\%}$ *ct* : *FrL*(q_4)	$\forall\exists$ *ct* : ($\top$)	65%	65%	100%
2	$\forall\exists$ *ct* : ($\top$)	$\cap_{\geq 40\%}$ *ct* : *FrL*(q_4)	100%	96%	96%
3	*numV(lo)*	$\cap_{\geq 60\%}$ *ct* : *FrL*(q_4)	35%	27%	79%
4	*numV(lo)*	*dens(hi)*	35%	29%	84%
5	*dens(hi)*	*numV(lo)*	32%	30%	93%
6	*dens(lo)*	*numV(hi)*	35%	31%	88%
7	*numV(hi)*	$\cap_{\geq 40\%}$ *ct* : *FrL*(q_4)	33%	31%	94%
8	*lSCC(lo)*	$\cap_{\geq 50\%}$ *ct* : *FrL*(q_4)	35%	35%	100%
9	*lSCC(lo)*	$\cap_{\geq 60\%}$ *ct* : *FrL*(q_4)	35%	27%	79%
10	*lSCC(lo)*	*numV(lo)*	35%	32%	94%
11	$\forall\exists$*in* : (*numV(hi)*)	*FrL(q1)*	37%	27%	76%
12	*FrL(q1)*	$\forall\exists$*in* : (*numV(hi)*)	47%	28%	59%
13	$\forall\exists$*in* : (*numV(lo)*)	*FrL(q1)*	6.5%	6.2%	94%
14	*FrL(q1)*	$\forall\exists$*in* : (*numV(lo)*)	47%	6%	13%

C. Results

We now present the result of our experiment. They are summarized by association rules extracted by RCA.

Based on RCA's results after one iteration, we discovered interesting rules on the words (24 954 rules) as well as on the dictionaries (206 476 rules). Some of these rules are presented in Table VII. The rules are indexed in the first column. Other columns are described at the beginning of subsection V-A.

It is worth mentioning that we focused our rules extraction on the those related to the size of the dictionary (*numV*). Rules found in Table VII can be interpreted as follows:

#1 65% of the dictionaries contain more than 60% of words frequently used by the players (FrL(q_4)).

#2 96% of the dictionaries contain more than 40% of frequently used words (FrL(q_4))

Already, those two first rules suggest that players tend to use a significant set of common words.

#3 79% of the small dictionaries (*numV(lo)*) contain more than 60% of frequently used words (FrL(q_4)).

#4 85% of small dictionaries (*numV(lo)*) are dense (*dens(hi)*).

#5 93% of dense (*dens(hi)*) dictionaries are small (*numV(lo)*).

On one hand, these three rules show that, among the dense dictionaries, the probability of having 60% of frequent word (*FrL*(q_4)) increase when compared to the same probability for all dictionaries (79% against 65%). Moreover, there is a strong correlation between small and dense dictionaries.

#6 88% of sparse dictionaries (*dens*(lo)) are large (*numV*(hi)) and contain more than 60% words frequently used (*FrL*(q_4)).

#7 94% of large dictionaries (*numV(hi)*) contain more than 40% of frequently used words (*FrL*(q_4)).

Conversely, large dictionaries are sparse. Moreover, since there is no rule of the form *numV(hi)* $\rightarrow \cap_{>50\%}$ *ct* : *FrL*(q_4) is produced, we can state that 94% of large dictionaries contain more than 40% , but less than 50% of frequent words.

#8 All dictionaries having a small largest SCC (*lSCC(lo)*) have at least half of their words frequently used (*FrL*(q_4)).

#9 79% of dictionaries having a small largest SCC (*lSCC(lo)*) have at least 60% of words frequently used (*FrL*(q_4)).

#10 94% of dictionaries having a small largest SCC (*lSCC(lo)*) are small (*numV(lo)*).

Roughly speaking, a small largest SCC signifies that used words are more frequent and smaller.

#11 76% of words used only in large dictionaries (*numV(hi)*) are unfrequent (*FrL(lo)*).

#12 59% of rarely used words (*FrL(lo)*) are exclusively used in large dictionaries (*numV*(hi)).

#13 94% of words used exclusively in small dictionaries (*numV(lo)*) are unfrequent.

#14 13% of words unfrequent (*FrL(lo)*) words are exclusively used in small dictionaries (*numV(lo)*).

Hence, words exclusive to large dictionaries are unfrequent, and so are those exclusive to small dictionaries. However, unfrequent words are more characteristic to large dictionaries.

VI. Discussion and Concluding Remarks

From those observations, it seems that the following latent scenario is followed. For a given seed word, there are ideal sets of words that should be chosen. These ideal sets form a dictionary by minimizing the number of words, by exploiting a stronger density and, in particular, the density of the largest strongly connected component. However, players sometimes have difficulty to formulate those more complex definitions, and then diverge from these ideal sets of words. Two tendencies seem to prevail. In most of the cases, when a player hesitates over a definition, this definition loses concision and several definitions must be produced to compensate. There is no special reason to expect these divergences on the same words for different players. Words exclusive to small dictionaries suggest the existence of another answer from the players when facing more complex definitions: the players fall back on synonymy (see Figure 1) to avoid providing a complete and unambiguous definition.

One surprising discovery we made in the experiment was the absence of significant association rules related to the two other psycholinguistic variables (age of acquisition and concreteness). Almost all extracted rules have either weak antecedent support or weak rule support. It is also important to mention some limits of our experiments. First, the variable *numV* seems to be a rough and convenient statistic to measure the quality of a strategy for the dictionary game. However, it does not take into account the concision, the precision and the pertinence of the definitions. Moreover, the seed word seems to play a significant role in the observations. Consequently, in a future experiment, we intend to normalize the frequencies with respect to their seed words. In the same spirit, we plan to verify if the seed word is, by itself, important, or if it is only its psycholinguistic category that impacts the dictionary structure.

References

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

[2] A. Blondin Massé et al., “How is meaning grounded in dictionary definitions?” in Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing, 2008.

[3] P. Vincent-Lamarre et al., “The latent structure of dictionaries,” Topics in cognitive science, 2016.

[4] M. Rouane-Hacene et al., “Relational concept analysis: mining concept lattices from multi-relational data,” AMAI, 2013.

[5] B. Ganter and R. Wille, Formal concept analysis: mathematical foundations, 1999.

[6] C. Goddard et al., “Introducing lexical primitives,” Semantic and Lexical Universals. Theory and Empirical Findings, 1994.

[7] M. Taddeo et al., “Solving the symbol grounding problem: a critical review of fifteen years of research,” Journal of Experimental & Theoretical AI, 2005.

[8] C. K. Ogden, “Basic English: A general introduction with rules and grammar, paul treber & co,” Ltd. London, vol. 1940, 1930.

[9] ——, “Ogden’s basic english,” 2018. [Online]. Available: http://ogden.basic-english.org/wordmenu.html

[10] C. Garrido et al., “Dictionaries as networks: Identifying the graph structure of ogden’s basic english,” in Proceedings of COLING, 2016.

[11] A. Wierzbicka, Semantic Primitives. (Frankfurt/M.)Athenäum-Verl., 1972.

[12] ——, Semantics: Primes and universals: Primes and universals. Oxford University Press, UK, 1996.

[13] C. Browne, “The new general service list: A core vocabulary for EFL students & teachers,” 2013.

[14] ——, “A new general service list: The better mousetrap we’ve been looking for,” Vocabulary Learning and Instruction, vol. 3, 2014.

[15] C. Browne et al., “The new general service list: Celebrating 60 years of vocabulary learning,” The Language Teacher, 2013.

[16] C. Goddard, Minimal English for a global world. Springer, 2018.

[17] A. Wierzbicka, “Natural semantic metalanguage,” 2020. [Online]. Available: https://intranet.secure.griffith.edu.au/schools-departments/natural-semantic-metalanguage/what-is-nsm

[18] J.-M. Poulin et al., “Strategies for learning lexemes efficiently: A graph-based approach,” in COGNITIVE 2018: The Tenth International Conference on Advanced Cognitive Technologies and Applications, 2018.

[19] R. Tarjan, “Depth-first search and linear graph algorithms,” SIAM journal on computing, vol. 1, no. 2, 1972, pp. 146–160.

[20] O. Picard et al., “Hidden structure and function in the lexicon,” arXiv preprint arXiv:1308.2428, 2013.

[21] P. Vincent-Lamarre et al., “The dictionary game,” 2017. [Online]. Available: http://lexis.uqam.ca:8080/dictGame/index.jsp

[22] V. Kuperman et al., “Age-of-acquisition ratings for 30,000 english words,” Behavior Research Methods, 2012.

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

[24] M. Brysbaert et al., “Concreteness ratings for 40 thousand generally known english word lemmas,” Behavior research methods, 2014.

[25] 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, 2009, pp. 977–990.

[26] R. Agrawal et al., “Fast discovery of association rules.” Advances in knowledge discovery and data mining, vol. 12, no. 1, 1996.

[27] M. Kryszkiewicz, “Concise Representations of Association Rules,” in Pattern Detection and Discovery. Springer, 2002, vol. 2447.

[28] S. Kramer et al., “Propositionalization approaches to relational data mining,” in Relational Data Mining, 2001, pp. 262–291.

[29] M. Wajnberg et al., “Concept analysis-based association mining from linked data: A case in industrial decision making,” JOWO, 2019.