An Optimal and Portable Parsing Method for Romanian, French, and German Large Dictionaries

Neculai Curteanu
Institute of Computer Science,
Romanian Academy,
Iaşi Branch
ncurteanu@yahoo.com

Alex Moruz
Institute of Computer Science, Romanian Academy;
Faculty of Computer Science,
"Al. I. Cuza" University, Iaşi
mmoruz@info.uaic.ro

Diana Trandabăţ
Institute for Computer Science, Romanian Academy;
Faculty of Computer Science,
"Al. I. Cuza" University, Iaşi
dtrandabat@info.uaic.ro

Abstract

This paper presents a cross-linguistic analysis of the largest dictionaries currently existing for Romanian, French, and German, and a new, robust and portable method for Dictionary Entry Parsing (DEP), based on Segmentation-Cohesion-Dependency (SCD) configurations. The SCD configurations are applied successively on each dictionary entry to identify its lexicographic segments (the first SCD configuration), to extract its sense tree (the second configuration), and to parse its atomic sense definitions (the third one). Using previous results on **DLR** (The Romanian Thesaurus – new format), the present paper adapts and applies the SCD-based technology to other four large and complex thesauri: **DAR** (The Romanian Thesaurus – old format), **TLF** (Le Trésor de la Langue Française), **DWB** (Deutsches Wörterbuch – GRIMM), and **GWB** (Göthe-Wörterbuch). This experiment is illustrated on significantly large parsed entries of these thesauri, and proved the following features: (1) the SCD-based method is a completely *formal grammar-free* approach for dictionary parsing, with efficient (weeks-time adaptable) modeling through sense hierarchies and parsing portability for a new dictionary. (2) SCD-configurations separate and run sequentially and independently the processes of lexicographic segment recognition, sense tree extraction, and atomic definition parsing. (3) The whole DEP process with SCD-configurations is *optimal*. (4) SCD-configurations, through sense marker classes and their dependency hypergraphs, offer an unique instrument of lexicon construction comparison, sense concept design and DEP standardization.

1 Introduction

The general idea behind parsing a large dictionary can be reduced to transforming a raw text entry into an indexable linguistic resource. Thus, for each dictionary entry, a structured representation of its senses has to be created, together with a detailed description of the entry's form: i.e. morphology, syntax, orthography, phonetics, lexical semantics, etymology, usage, variants etc.

The aim of this paper is to present an efficient *dictionary entry parsing* (DEP) method, based on Segmentation-Cohesion-Dependency (SCD) configurations (Curteanu, 2006), applied on a set of five large and complex dictionaries: **DLR** (The Romanian Thesaurus – new format), **DAR** (The Romanian Thesaurus – old format), **TLF** (Le Trésor de la Langue Française), **DWB** (Deutsches Wörterbuch – GRIMM), and **GWB** (Göthe-Wörterbuch).

The paper is structured in 8 sections: Section 2 presents the state of the art in DEP, with an emphasis on the comparison between the proposed method and other dictionary parsing strategies, before detailing the SCD-based proposed method in Section 3. The following sections present the application of the proposed method to the five dictionaries. The paper ends with a discussion on

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), pages 38–47,
Beijing, August 2010

comparative results and development directions concerning optimality, portability, standardization, and dictionary networks.

2 Dictionary Entry Parsing

Natural language text parsing is a complex process whose prerequisite essential stage is a thorough modeling of the linguistic process to be developed, *i.e.* the structures and relations aimed to constitute the final result of the analysis. Similarly, for DEP, the semantics of the lexical structures, the sense markers, and the hierarchies (dependencies) between sense structures must be specified.

Standard approaches to dictionary entry parsing (referred to from now on as *standard* DEP), such as the one used by (Neff and Boguraev, 1989), the *LexParse* system presented in (Hauser and Storrer, 1993; Kammerer, 2000; Lemnitzer and Kunze, 2005), or lexicographic grammars, as those presented in (Curteanu & Amihaesei, 2004; Tufis et al., 1999), recognize the sense / subsense definitions in a strictly sequential manner, along with the incremental building of the entry sense tree. The interleaving of the two running processes is the main source of errors and inefficiency for the whole DEP process.

Both the *standard* DEP (Figure 1) and our proposed method based on SCD-*configurations* (Figure 2) involve the following *three* running *cycles* and *four* essential *phases* for extracting the sense-tree structure from a dictionary:

[A1], **[B1]** – parsing the *lexicographic segments* of an entry;

[A2], **[B2]** – parsing the *sense-description segment* of the dictionary entry, at the level of explicitly defined senses, until and not including the contents of the atomic definitions / senses; at this stage, the *sense-tree* of the sense-description segment is built having (sets of) atomic senses / definitions in their leaf-nodes.

[A3], **[B3]** – parsing the *atomic definitions* / *senses*.

Phase_1 := Sense-*i* Marker Recognition;
Phase_2 := Sense-*i* Definition Parsing;
Phase_3 := Attach Parsed Sense-*i* Definition to Node-*i*;
Phase_4 := Add Node-*i* to EntrySense-Tree.

The parsing cycles and phases of existing approaches, called *standard* DEP, are summarized by the pseudo-code in Fig. 1, where *MarkerNumber* is the number of markers in the dictionary-entry marker sequence and *EntrySegmentNumber* is the number of lexicographic segments of the parsed entry.

[A1]. For *s* from 1 to *EntrySegmentNumber*
If(**Segment-*s* = Sense-Segment**)
[A2]. For *i* from 0 to *MarkerNumber*
Phase_1 Sense-*i* Marker Recognition;
Phase_2 Sense-*i* Definition Parsing;
[A3]. *If*(Success)
Phase_3 Attach Parsed Sense-*i* Defi nition to Node-*i*;
Phase_4 Add Node-*i* to **Entry Sense Tree;**
[/A3]. *Else* Fail and Stop.
[/A2]. EndFor
Output: EntrySenseTree *with* **Parsed Sense Definitions**
(only if **all** sense definitions are **parsed**).
Else **Segment-*s* Parsing;**
Continue
[/A1]. EndFor
Output: Entry parsed segments, including the Sense-Segment (only if all definitions in the Sense-Segment are parsed).

Fig. 1. Standard dictionary entry parsing

The main drawback of the classical, *standard* DEP, is the embedding of the parsing cycles, [A1] [A2] [A3] ... [/A3] [/A2] [/A1], derived from the intuitive, but highly inefficient parsing strategy based on the general Depth-First searching. After presenting the SCD-based dictionary parsing method, section 3.2. compares the parsing cycles and phases of standard DEP to the ones of SCD-based DEP.

3 Parsing with SCD Configurations

The SCD *configuration*(*s*) method is a procedural, recognition-generation computational device, that is distinct from the traditional and cumbersome *formal grammars*, being able to successfully replace them for several tasks of natural language parsing, including text free parsing (Curteanu, 2006) and thesauri parsing (Curteanu et al., 2008). For SCD-based parsing, the semantics and the linguistic modeling of the text to be analyzed should be clearly specified at each parsing level, and implemented within the following components of each SCD configuration (hereafter, SCD-*config*):

- A set of *marker classes*: a *marker* is a boundary for a specific linguistic category (*e.g.* **A.**, **I.**, **1.**, **a).**, etc.). Markers are joined into *marker classes*, with respect to

their functional similarity (*e.g.* {**A.**, **B.**, **C.**, …}, {**1.**, **2.**, **3.**, …}, {**a).**, **b).**, …});

- A *hypergraph-like hierarchy* that establishes the dependencies among the marker classes;
- A *searching* (*parsing*) *algorithm*.

Once an SCD configuration is defined, parsing with the SCD configuration implies identifying the markers in the text to be parsed, constructing the *sequences* of markers and categories, recognizing the marked text structures (spans within the bounding markers) corresponding to the SCD configuration semantics, and classifying them according to the marker sequences within the pre-established hierarchy assigned to that SCD configuration. The last step settles the dependencies and correlations among the parsed textual structures. Identifying the lexicographic segments of an entry, the syntactic and semantic structure of each segment, the senses, definitions and their corresponding markers, is the result of an in-depth lexical semantics analysis. Designing the classes and the hypergraph structure of their dependencies are essential cognitive aspects of working with SCD configurations, and need to be pre-established for each dictionary.

Within the parsing process, each SCD *configuration*, *i.e.* marker classes, hierarchy, and searching algorithm, is completely commanded by its *attached* semantics. The semantically-driven parsing process, either for free or specialized texts, consists in a number of SCD configurations applied sequentially (in cascade), each one on a different semantic level. The semantic levels (each one driving an SCD configuration) are *subsuming* each other in a top-down, monotonic manner, starting with the most general semantics of the largest text span, until the most specific level.

3.1 SCD Configurations for DEP

The SCD-based process for DEP consists in three SCD *configurations*, applied sequentially on the levels and sublevels of the dictionary entry, where each level should be monotonic at the lexical-semantics *subsumption* relation.

The task of applying the SCD *configurations* to DEP requires knowing the semantics of the corresponding classes of sense and definition markers, together with their hierarchical representation.

The first SCD configuration (**SCD-*config*1**) is devoted to the task of obtaining the partition of the entry *lexicographic segments* (Hauser & Storrer, 1993). Since usually there are no dependency relations between lexicographic segments, SCD-*config*1 is not planned to establish the dependency relations (cycle **[A1]** in Fig. 1, or cycle **[B1]** in Fig. 2).

The *second* important *task* of DEP is to parse each lexicographic segment according to its specific semantics and linguistic structure. The most prominent lexicographic segment of each entry is the *sense-description one*, the central operation being the process of extracting the *sense tree* from the segment. This is the purpose of the *second* SCD *configuration* (denoted **SCD-*config*2**), corresponding exactly to the DSSD parsing algorithm in (Curteanu et al., 2008), which, for the **DLR** sense trees, has a precision of 91.18%. In order to refine the lexical-semantics of primary senses, one has to descend, under secondary senses, into the definitions and definition examples, which constitute the text spans situated between two sequentially-related nodes of the parsed sense tree. This SCD configuration is represented as cycle **[B2]** in Fig. 2.

The *third step* of DEP parsing (cycle **[B3]** in Fig. 2) is represented by the configuration **SCD-*config*3**, needed to complete the DEP. SCD-*config*3 consists in a specific set of marker classes for the *segmentation* at dictionary definitions, the hypergraph-like hierarchy of the classes of markers for these sense definitions, and the parsing algorithm to establish the dependencies among atomic senses / definitions. As a prerequisite groundwork, an *adequate modeling* of the sense definitions is needed and *the segmentation* of *definitions* is implemented as an essential step to establish the dependency-by-subsumption among the sense types of the considered thesaurus. The final result of the entry parsing process should be the sequential application of the SCD-*config*1, SCD-*config*2, and SCD-*config*3 configurations.

3.2 A Structural Analysis: Standard DEP *vs.* SCD Configurations

A pilot experiment of parsing with SCD configurations was its application to the **DLR** thesaurus parsing (Curteanu et al., 2008); the process of *sense tree building* has been completely detached

and extracted from the process of *sense definition parsing*.

The sense-tree parsing with *SCD-based* DEP cycles and phases is summarized in pseudo-code in Fig. 2 and comparative Table 1 below.

[B1]. For *s* from 1 to *EntrySegmentNumber*
Segment-*s* Parsing;
If(**Segment-*s* = Sense-Segment**)
Standby on **Sense-Segment Parsing;**
Else Continue
[/B1]. EndFor
Output: Entry parsed segments, not including the Sense-Segment;
[B2]. For *i* from 0 to *MarkerNumber*
Phase_1 Sense-*i* Marker Recognition;
Assign (Unparsed) Sense-*i* Definition to **Node-*i*;**
Phase_4 Add Node-*i* to **EntrySenseTree;**
Standby on **Sense-*i* Definition Parsing;**
[/B2]. EndFor
Output: EntrySenseTree (with **unparsed** sense definitions)**.**
Node-*k* = Root(**EntrySenseTree**);
[B3]. *While* not all nodes in **EntrySenseTree** are visited
Phase_2 Sense-*k* Definition Parsing;
If(Success)
Phase_3 Attach Parsed Sense-*k* Definition to Node-*k*;
Else **Attach Sense-*k* Parsing Result to Node-*k*;**
Node-*k* = getNextDepth FirstNode(**EntrySenseTree**)
Continue
[/B3]. EndWhile.
Output: EntrySenseTree (with **parsed** or **unparsed** definitions)**.**
Output: Entry parsed segments, including the Sense-Segment.

Fig. 2. SCD-based dictionary entry parsing

Standard DEP	*SCD-based DEP*
(Phase_1; **Phase_2** **(Phase_3;** **Phase_4)** **)**	**(Phase_1;** **Phase_4)** **(Phase_2;** **Phase_3)**

Table 1: Dictionary parsing phases in standard DEP and SCD-based DEP

Table 1 presents the ordering of the dictionary parsing phases in the *standard* DEP strategy (the four phases are embedded) and the SCD-based DEP strategy (the phases are organized in a linearly sequential order).

Since the process of *sense tree construction* (cycle **Phase_1** + **Phase_4**) has been made completely detachable from the *parsing* of *the* (atomic) *sense definitions* (cycle **Phase_2** + **Phase_3**), the whole SCD-based DEP process is much more efficient and robust. An efficiency *feature* of the SCD-based parsing technique is that, working exclusively on sense marker sequences, outputs of **[B2]** and **[B3]** cycles in Fig. 2 (*i.e.* sense trees) are obtained either the sense definition parsing process succeeds or not, either correct or not!

These properties of the new parsing method with SCD configurations have been effectively supported by the parsing experiments on large Romanian, French, and German dictionaries.

4 Romanian DLR Parsing

The study of the application of SCD-configuration to DEP started with the analysis of the **DLR** parsing (Curteanu et al., 2008). Fig. 3 presents the hierarchy of SCD-*config*2 for **DLR** sense marker classes,

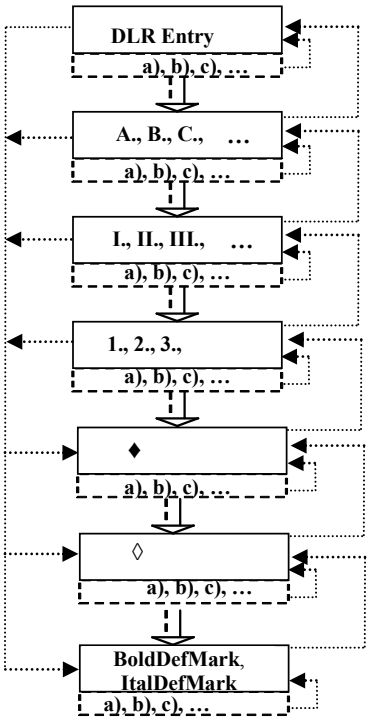


Fig. 3. Hierarchy of DLR marker classes

devoted to *sense tree parsing*. The *dashed arrows* point to the upper or lower levels of **DLR** sense marker hierarchy, from the *literal enumeration* layer-embedded level. The *continuous-dashed arrows* in Fig. 3 point downwards from the higher to the lower priority levels of **DLR** marker class hypergraph. Because of its special representation characteristics, the literal enumeration is illustrated on a layer *attached* to the hie-

rarchy level (dashed line) to which it belongs, on *each* of the sense levels.

A detailed description of the **DLR** sense tree extraction with SCD-*config*2 (denoted as DSSD algorithm) is found in (Curteanu et al., 2008).

4.1 DLR Parsing: Problems and Results

The three SCD-*configurations* establish the dependencies among **DLR** senses (SCD-*config*1-2) and definitions (SCD-*config*3). However, **DLR** is encoded by default *inheritance* rules of senses (definitions), acting on all the node levels of the sense / definition trees.

The sense tree parser (output of SCD-*config*2) was tested on more than 500 dictionary entries of large and medium sizes. The success rate was 91.18%, being computed as a perfect match between the output of the program and the gold standard. Furthermore, it is worth noting that an entry with only one incorrect parse (*i.e.* one node in the sense tree attached incorrectly) was considered to be erroneously parsed in its entirety, an approach which disregards all the other correctly attached nodes in that entry.

A first source of parsing errors is the non-monotony of the marker values: "**A.** [**B.** missing] … **C.** …"; "**2.** [instead of **1.**]... **2.** …"; "…**a)**… **b)** … **c)** … **b)** [instead of **d)**] …". Another major source of parsing errors comes from the inherent ambiguity in deciding which is the regent and which is the dependent (sub)sense in marker sequences as "**1. a) b) c)** ◊ [◊]…".

For evaluating SCD-*config*3, 52 dictionary entries of various sizes were used as a gold standard, totaling a number of approximately 2000 chunks and 22,000 words. The results are given in Table 2. Upon further analysis of the evaluation results, the most frequent errors were found to be due to faulty *sigle* (abbreviation of the source of examples) segmentation. A detailed analysis of the error types for the DLR dictionary is discussed in (Curteanu et al., 2009).

Evaluation Type	Precision	Recall	F-measure
Exact Match	84.32%	72.09%	77.73%
Overlap	92.18%	91.97%	92.07%

Table 2: Evaluation results for segmentation of DLR atomic sense elements

Correcting the acquisition of *sigles* leads to a 94.43% *f-measure* for *exact match* (the number of correctly identified sense and definition units) and a 98.01% *f-measure* for *overlap* (the number of correctly classified words in each entry). To achieve the **DLR** parsing process completely, the last operation to be executed is to establish the dependency relations between atomic senses / definitions, under all the sense nodes in the computed sense-tree of the entry. Currently, the **DLR** is parsed almost completely, including at atomic senses / definitions, the lexicographic segments and sense-trees being obtained with a correctness rate above 90% for explicitly marked sense definitions.

5 Romanian DAR Parsing

The structure of the main lexicographical segments in **DAR** is outlined below:

I. The *French Translation* segment, denoted *FreSeg*, contains the French translations of the lemma and the main sense hierarchy of the entry word. The translation of the sense structure into Romanian and the complete description of the sense tree in **DAR** are in a subsumption relation. In some cases, the French translation may not exist for specific Romanian lemmas.

II. The *general description* segment (*RomSeg*) is composed of several paragraphs and contains morphologic, syntactic, semantic, or usage information on the entry word. *RomSeg* usually starts with the entry word in italics (otherwise, the entry word occurs in the first row of the first paragraph).

III. The third segment of a **DAR** entry, called *SenseSeg*, is the *lexical-semantic description* of the entry word. *SenseSeg* is the main objective of the lexicographic analysis of the entry parsing in order to obtain its sense tree.

IV. The fourth segment of a **DAR**, *NestSeg*, contains one or more "*nests*", which are segments of text describing morphological, syntactic, phonological, regional, etc. variants of an entry, sometimes with an attached description of the specific senses. The structure of the **DAR** *nest* segment is similar to that of a typical **DAR** entry, and the recursive nature of **DAR** entries comes from the sense parsing of *nest* segments.

V. The fifth segment of **DAR** entries, denoted *EtymSeg*, contains etymological descriptions of the entry word and is introduced by an *etymology-dash* (long dash "–"). Among the five segments of a **DAR** entry, the only compulsory ones are *FreSeg* and *SenseSeg*. The other three segments are optional in the entry description, depending on each entry word.

5.1 DAR Marker Classes and Hierarchy

The priority ordering of **DAR** marker classes is:
1. Capital letters (*LatCapLett_Enum*): A., B., ...
2. Capital roman numerals (*LatCapNumb_Enum*): I., II., ...
3. Arabic numerals (*ArabNumb_Enum*): 1^0., 2^0. These markers introduce the *primary senses*, in a similar manner to those in **DLR**.
4. For introducing *secondary senses*, **DAR** uses the same sense markers used in **DLR** for definitions of type *MorfDef*, *RegDef*, *BoldDef*, *ItalDef*, *SpecDef*, *SpSpecDef*, and *DefExem*, and a set of markers specific to **DAR**: ||, |, #, †.
5. According to the level of the lexical-semantic description, **DAR** uses *literal enumeration* on *two levels*: **(5.a)** lowercase Latin letters (*LatSmallLett_Enum*): a.), b.), … **(5.b)** a *LatSmallLett_Enum* can have another enumeration, using lowercase Greek letters (*GreSmallLett_Enum*): α.), β.), γ.), …

The hierarchies for sense markers in DAR are given in Fig. 4.

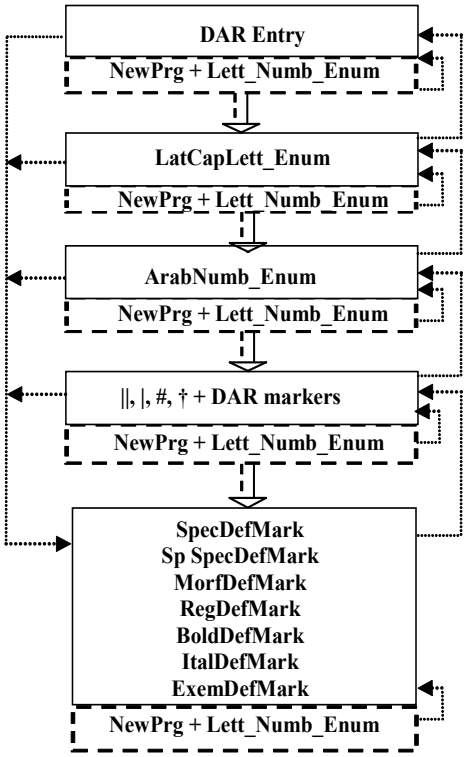
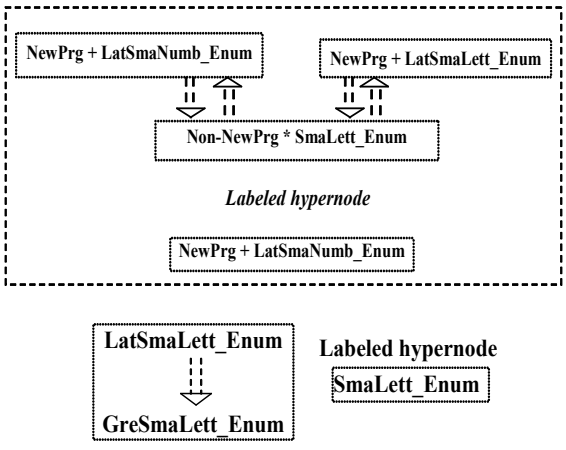


Fig. 4: Dependency hypergraph for DAR

5.2 Special problems in DAR parsing

A first difficulty for parsing the **DAR** lexicographic segments is the occurrence of the *New_Paragraph* (*NewPrg*). For *NewPrg* marker recognition we used an *implicit level of enumeration*, *LatSmaNum_Enum* discriminating the cases when *NewPrg* is (or not) accompanied by another type of **DAR** sense marker.



The second difficult problem in **DAR** parsing is the process of refining the *NewPrg* marker with literal enumerations (*LatSmallLett_Enum*), which can be in turn refined by an implicit enumeration using *NewPrg*. This has been solved by interpreting the sense levels for enumerations according to their context.

Using SCD configurations, we have parsed 37 **DAR** entries of large and medium sizes. The results of the **DAR** parsing have been evaluated manually, and we estimated the parsing precision as really promising, taking into account the difficult problems raised by **DAR**. The lack of a gold standard for **DAR** entries did not allow performing an automatic evaluation at this moment.

DAR Entry Parsing (Excerpt):

```
- <entry>
- <lexsegm value="FreSeg." class="0">
- <sense value="LARG, -Ă" class="1">
  <definition>adj., s. a. şi f. I. 1°. {i}Large, vaste.{/i} 2°. (Fig.)
{i}Large, ample, majestueux. Largement. {/i}3°. {i}Au large, à
…{i}Femme légère, dessalée{/i}.</definition>
- </sense >
- </lexsegm >
```

```
- <segm value="SenseSeg." class="0">
- <sense value="I." class="8">
- <definition>  A d j. şi a d v. </definition>
- <sense value="1°." class="12">
- <definition>
  A d j. (În opoziţie cu î n g u s t) Extins în toate direcţiile;
…{i}Larg {/i}{i}= {/i}largus.
  <SRCCITE source="ANON. CAR.">ANON.
CAR.</SRCCITE>  {i}Calea ceaia largă.{/i}
  <AUTHCITE source="EV." author="CORESI" sigla="CORESI,
EV." …
  </definition>
</sense>
- <sense value="2°." class="12">
- <definition>  F i g. (Literar, după fr.) Mare, amplu, …
  <AUTHCITE source="C. I." volume="II" …</AUTHCITE> …
  </definition>
- <sense value="‖" class="20">
  <definition>Adv. {i}Musafirul... se urca …</definition>
  </sense>
  </sense>
- <sense value="3°." class="12">
- <definition>  (În funcţiune predicativă,…)
…  …  …
</definition>
  </sense>
  </sense>
- <sense value="II." class="8">
- <definition>  S u b s t. </definition>
- <sense value="1°. " class="12">
- <definition>  S. a. Lărgime. {b}Inimii închise...  </definition>
  </sense>
…  …  …
- <sense value="2°." class="12">
- <sense value=" NewPrg " class="13">
  <definition>{i}LĂRGÍME{/i} s f. v. {b}larg{/b}.</definition>
  </sense>
- <sense value="NewPrg" class="13">
  <definition>{i}LĂRGAMĂNT{/i} † S. A. V.
{i}larg{/i}.</definition>
…  …  …
  </sense>
  </sense>
  </lexsegm>
  </entry>
```


6 French TLF Parsing

The French **TLF**, a very well-organized and structured thesaurus, provides both similarities and distinctive characteristics in comparison to the Romanian **DLR**. The structure of **TLF** lexicographic segments, obtained with the SCD-*config*1, is relatively simple. A **TLF** entry commences with the *sense-description segment*, optionally (but very frequently) followed by a package of "*final*" *segments*, introduced by specific labels, as in the pattern:

REM. 1. … **2.** … **3.** …
PRONONC. ET ORTH. – … Eng.: …
ÉTYMOL. ET HIST. I. ... **1. a)** … **b)** … **2.** …
3. … **II.** …
STAT. Fréq. abs. littér.: … **Fréq. rel. littér.:** …
DÉR. 1. …**2.** …**3. a)** … **b)** … **Rem. a)** … **b)** …
BBG. …

As one may notice, some *final segments* can enclose particular sense descriptions, similarly to those met in the proper sense-description segment. The sense markers in **TLF** resemble to those in **DLR**, but there are also significant differences. The dependency hypergraph of the **TLF** marker classes is the following:

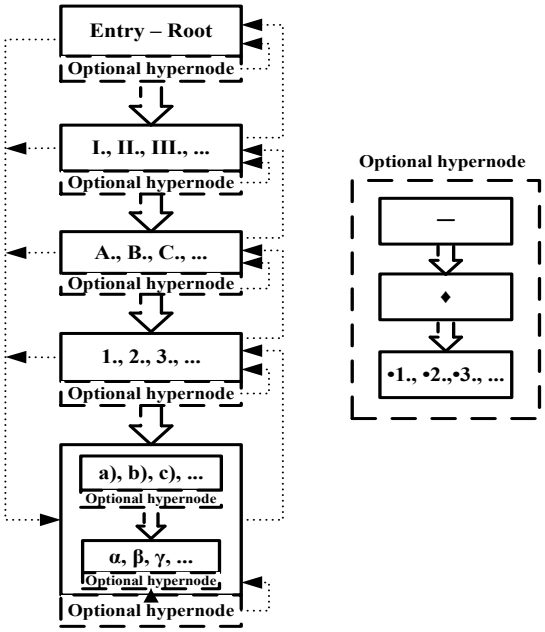


Fig. 5. Dependency hypergraph of TLF sense marker classes

Cross-linguistic hints involving **TLF** entry parsing with SCD configurations: (*a*) A new sense marker (compared to **DLR**) is "–" (*inheritance- dash*), aiming to signal the presence of an inherited sense. (*b*) When "–" occurs *after* another **TLF** marker, the "–" role is to inherit a parent sense (either regent or not) from the sense-tree. (*c*) When "–" begins at new paragraph (*NewPrg*), its role is of a intermediary subsense, inheriting the meaning of the regent (sub)sense. (*d*) Another new **TLF** marker is "•1., •2., …" (indexed, small red-ball), defining the *new* **TLF** sense concept: *Indexed Examples to Definitions* for the whole entry (denoted *IdxDefExem*). (*e*) The literal enumeration with Latin small letters (*LatSmaLett_Enum*) is refined with Greek small letters (*GreSmaLett_Enum*). (*f*) In **TLF**, only the *filled diamond* "♦" marker is present (as secondary sense); the *empty diamond* "◊" is missing. (*g*) Some primary senses ("**I.**", "**A.**") in

TLF receive reversed priorities (Fig. 5) in the marker class hierarchy as compared to **DLR**.

6.1 TLF Parsing Results

For **TLF**, we processed 31 significant entries (TLFi, 2010) of medium and large dimensions (entries of 5 pages each, A4 format, in average) with the parser based on SCD configurations. The parsing results have been evaluated manually, the correctness rate being above 90%. One of the reasons of these very promising results for **TLF** parsing may be the regularity and standardization of the **TLF** entry texts. An automatic, precise evaluation of the SCD-based parser was not possible since we are missing currently a gold-corpus of **TLF** entries.

TLF Entry Parsing (Excerpt):

```
- <entry>
- <lexsegm value="SenseSeg." class="0">
- <sense value="ANNONCER" class="1">
+ <definition> - <sense value="I." class="2">
- <definition>  <i>Emploi trans.</i> … …
  </definition>
- <sense value="A." class="3">
  <definition>[Le suj. désigne une pers.]</definition>
- <sense value="1." class="4">
- <definition>
  [L'obj. désigne un événement] Faire connaître …
  </definition>
- <sense value="a)" class="5">
- <definition>
  [L'événement concerne la vie quotidienne] …
  <i>Annoncer qqc. à qqn, annoncer une bonne</i> … …
  </definition>
- <sense value="circle" class="10">
- <definition>
  1. À la mi-novembre, Costals …
  <b>annonça</b>  son retour pour le 25. Dans la lettre …
  </definition>
  </sense>
- <sense value="circle" class="10">
- <definition>
  2. Électre, fille d'un père puissant, réduite…
  <b>annonce</b> … … …
  </definition>
  </sense>
  </sense>
- <sense value="b)" class="5">
- <definition>
  <i>JEUX (de cartes). Faire une annonce.</i> …
  </definition>
- <sense value="circle" class="10">
- <definition>
  3. Celui qui la détient la belote …
  <b>annonce</b>  alors :  <i>belote,</i>
… …
  </definition>
  </sense>
  </sense>
… … …
  </lexsegm>
- <lexsegm value="FinSeg." class="0">
- <sense value="-" class="5">
- <definition>  <b>ÉTYMOL. ET HIST.</b> … …
  <i>Ca</i>  1080  <i>anuncier</i>
… … …
  </definition>
  </sense>
…. … …
- <definition>
  <b>BBG.</b>  ALLMEN 1956. BRUANT 1901. …
  <b>ARRIVÉE, subst. fém.</b>
  </definition>
… … …
  </sense>
  </lexsegm>
  </entry>
```

7 Lexicographic Segments and Sense Markers in German DWB and GWB

The German **DWB** entries comprise a complex structure of the lexicographic segments, which provide a non-uniform and non-unitary composition (Das Woerterbuch-Netz, 2010). One special feature is that **DWB** and **GWB** lexicographic segments are composed of two parts: a first (optional) *root-sense* subsegment, and the segment *body*, which contain the explicit sense markers, easily recognizable. For **DWB**, the parsing of lexicographic segments is not at all a comfortable task since they are defined by three distinct means:

(A) *After* the *root-sense* of a **DWB** entry, or *after* the *root-sense* of a lexicographic segment, (a list of) italicized-and-spaced key-words are placed to constitute the *label* of the lexicographic segment *that follows*. Samples of such key-word labels for **DWB** lexicographic segments are: "*Form, Ausbildung und Ursprung*", "*Formen*", "*Ableitungen*", "*Verwandtschaft*","*Verwandtschaft und Form*", "*Formelles und Etymologisches*", "*Gebrauch*", "*Herkunft*", "*Grammatisches*", etc., or, for **DWB** sense-description segment: "*Bedeutung und Gebrauch*" (or just "*Bedeutung*"). In the example below, they are marked in grey.

GRUND, *m., dialektisch auch f. gemeingerm. wort; fraglich ist … poln. russ. slov. nlaus.* grunt *m.* *f o r m u n d h e r k u n f t* .

1) *für das verständnis der vorgeschichte des wortes ist die z w i e g e s c h l e c h t i g k e i t …*

H. V. SACHSENHEIM *spiegel* 177, 30;

städtechron. 3, 51, 14. … … … drey starcke grund 6, 290. *b e d e u t u n g* . *die bedeutungsgeschichte des wortes … …*

I. grund *bezeichnet die feste untere begrenzung eines dinges.*

A. *grund von gewässern; seit ältester zeit belegbar: profundum* (*sc. mare*) crunt *ahd. gl.* 1, 232, 18;

1) *am häufigsten vom meer* (*in übereinstimmung mit dem anord. gebrauch*): *...*

(B) *The second* way to specify the **DWB** current lexicographic segments is to use their labels as key-words immediately *after* the primary sense markers.

(C) The *third* (and most frequent) way to identify the lexical description segment(s) of a **DWB** entry is simply the lack of a segment label at the beginning of the sense description segment. By default, after the entry *root-sense segment* (which can be reduced to the Latin translation of the German lemma) the sense-description segment comes, without any "*Bedeutung*" label, introducing explicit sense markers and definitions.

7.1 German DWB and GWB Dependency Hypergraphs. Parsing Results

Without coming into details (see the marker class dependency hypergraphs in Fig.6 and Fig.7), one can say with a good measure of truth that a general resemblance hold between **DAR**

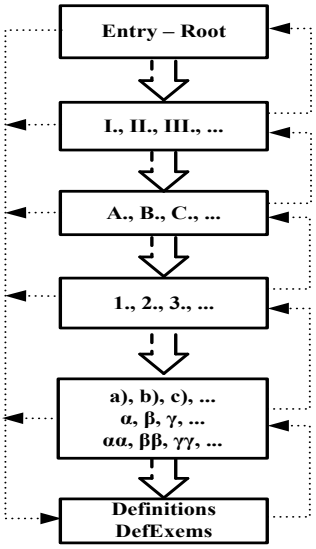


Fig. 6. DWB dependency hypergraph

and **DWB**, and **TLF** and **GWB**, respectively. The sense markers in **DWB** are usual, with the remark that sense refinement by literal enumeration is realized on three levels: *LatSmaLett_Enum* (**a)**, **b)**, …), *GreSmaLett_Enum* (**α)**, **β)**, …), and *GreDoubleSmaLett_Enum* (**αα)**, **ββ)**, …).

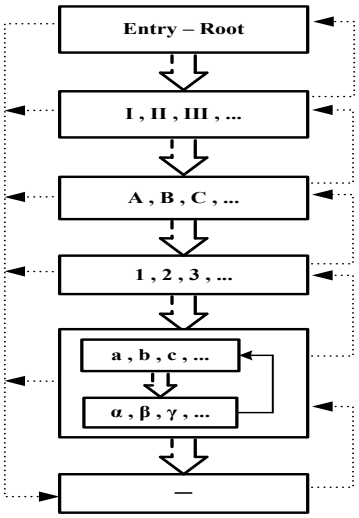


Fig. 7. GWB dependency hypergraph

A number of 17 very large **DWB** entries have been parsed only with SCD-*config*1 and SCD-*config*2. We appreciate on this small but significant excerpt of **DWB** entries that parsing of the sense description segment at sense trees is performed with a high precision, but delimitation of the lexicographic (sub-)segments and labels is a more difficult problem. The lack of a **DWB** entry gold corpus did not allow a precise, automated evaluation of the parser.

8 Directions: Optimal Lexicon Design, Standardization, Lexicon Networks

The special features of parsing with **SCD configurations** (SCD-*configs*) are: • SCD-configs is a completely *formal grammar-free* approach which involves simple, efficient (weeks-time adaptable), thus portable modeling and programs. • In all currently existing DEP methods, the sense tree construction of each entry is, more or less, recursively embedded and mixed within the definition parsing procedures. • SCD-*configs* provides a lexical-semantics refinement level on each SCD-*config*. • SCD-*configs* separate and run sequentially, on independent levels (*viz.* configurations), the processes of lexicographic segment recognition, sense tree extraction, and atomic definition parsing. • This makes the whole DEP process with SCD-*configs* to be *optimal*. • The sense marker classes and their dependency hypergraphs, specific to each thesaurus, offer an unique instrument of lexicon con-

struction comparison, sense concept design and standardization. With the SCD parsing technique, one can easily compare the sense categories, their marking devices, the complexity and recursiveness measure of the sense dependency hypergraphs for each thesaurus.

The cross-linguistic analysis of the five large thesauri showed the necessity of a careful lexical-semantics modeling of each dictionary. Equally important, many semantic and lexicographic concepts such as sense markers and definitions, (indexed) examples to definitions, sense and source references etc. can be similar, adaptable, and transferable between corresponding SCD-configurations of different thesauri.

The SCD-*configs* analysis pointed out the need of a more general and adequate terminology for the lexical-semantics notions. *E.g.*, comparing the Romanian and French thesauri with the German ones, we decided that, while preserving the definition type labels *MorfDef*, *DefExem*, *SpecDef* and *SpSpecDef*, we should change the *RegDef* into *GlossDef*, *BoldDef* into *IdiomDef*, *ItalDef* into *CollocDef*, and add the **TLF** *IdxDefExem* (an indexed *DefExem*) to the sense concept set.

The future experiments will continue with new thesauri parsing: Russian, Spanish, Italian, but the true challenge shall be oriented towards Chinese / Japanese thesauri, aiming to establish a thorough lexical-semantics comparison and a language-independent, portable DEP technology based on SCD configurations. A further development would be to align the Romanian thesauri sense and definition types to TEI P5 standards (XCES, 2007), and to design an optimal and cross-linguistic compatible *network of Romanian electronic dictionaries*, similar to a very good project of dictionary network, *i.e.* the German Woerterbuch-Netz (with links to **TLFi** entries too), whose twelve component lexicons include **DWB** and **GWB.**

Acknowledgement. The present research was partly financed within the **eDTLR** grant, PNCDI II Project No. 91_013/18.09.2007.

References

DLR revision committee. (1952). *Coding rules for DLR* (in Romanian). Romanian Academy, Institute of Philology, Bucharest.

Cristea, D., Răschip, M., Forăscu, C., Haja, G., Florescu, C., Aldea, B., Dănilă, E. (2007). *The Digital Form of the Thesaurus Dictionary of the Romanian Language*. In Proc. of the 4th SpeD 2007.

Curteanu, N., and E. Amihăesei. (2004). *Grammar-based Java Parsers for DEX and DTLR Romanian Dictionaries*. ECIT-2004, Iasi, Romania.

Curteanu, N. (2006). *Local and Global Parsing with Functional* (F)X-*bar Theory and SCD Linguistic Strategy*. (I.+II.), Computer Science Journal of Moldova, Academy of Science of Moldova, Vol. 14 no. 1 (40):74-102; no. 2 (41):155-182.

Curteanu, N., D. Trandabăţ, A. M. Moruz. (2008). *Extracting Sense Trees from the Romanian Thesaurus by Sense Segmentation & Dependency Parsing*, Proceedings of CogAlex Workshop, COLING 2008, ISBN 978-1-905593-56-9, :55-63.

Curteanu, N., Moruz, A., Trandabăţ, D., Bolea, C., Spătaru, M., Husarciuc, M. (2009). *Sense tree parsing and definition segmentation in eDTLR Thesaurus*, in Trandabăţ et al. (Eds.), Proc. of the Workshop "Linguistic Resources and Instruments for Romanian Language Processing", Iasi, Romania, "Al.I.Cuza" University Publishing House, ISSN 1843-911X, pp. 65-74, (in Romanian).

Das Woerterbuch-Netz (2010): http://germazope.uni-trier.de/Projects/WBB/woerterbuecher/

Hauser, R., and A. Storrer. (1993). *Dictionary Entry Parsing Using the LexParse System*. Lexikographica (9): 174-219.

Kammerer, M. (2000). *Wöterbuchparsing Grundsätzliche Überlegungen und ein Kurzbericht über praktische Erfahrungen*, http://www.matthias-kammerer.de/content/WBParsing.pdf

Le Trésor de la Langue Française informatisé (2010). http://atilf.atilf.fr/tlf.htm

Lemnitzer, L., and C. Kunze. (2005). *Dictionary Entry Parsing*, ESSLLI 2005.

Neff, M., and B. Boguraev. (1989). *Dictionaries, Dictionary Grammars and Dictionary Entry Parsing*, Proc. of the 27th ACL Vancouver, British Columbia, Canada, :91 – 101.

Tufiş, Dan. (2001). From Machine Readable Dictionaries to Lexical Databases, RACAI, Romanian Academy, Bucharest, Romania.

XCES TEI Standard, Variant P5. (2007). http://www.tei-c.org/Guidelines/P5/