ParaDis and Démonette: From Theory to Resources for Derivational Paradigms

Fiammetta Namer, Nabil Hathout

ParaDis and Démonette
From Theory to Resources for Derivational Paradigms

Fiammetta Namer
UMR 7118 ATILF
CNRS & Université de Lorraine
Nancy, France
fiammetta.namer@univ-lorraine.fr

Nabil Hathout
UMR 5263 CLLE-ERSS - CNRS &
Université de Toulouse Jean-Jaurès
Toulouse, France
nabil.hathout@univ-tlse.fr

Abstract

This article traces the genesis of the French derivational database Démonette$_{v2}$ and shows how current architecture and content of derivational morphology resources result from theoretical developments in derivational morphology and from the users' need. The development of this large-scale resource began a year ago and is part of the Demonext project (ANR-17-CE23-0005). Its conception is adapted from theoretical approaches of derivational morphology where lexemes, units of analysis, are grouped into families that are organized into paradigms. More precisely, Démonette$_{v2}$ is basically an implementation of ParaDis, a paradigmatic model for representing morphologically complex lexical units, formed by regular processes or presenting discrepancies between form and meaning. The article focuses on the principles of morphological, structural and semantic encoding that reflect the methodological choices that have been made in Démonette$_{v2}$. Our proposal will be illustrated with various examples of non-canonical word formations.

1 Introduction

Morphological analysis is one of the initial steps in many NLP systems. Analyzers, most often based on machine learning and statistical methods, decompose words into morphemes in order to compensate for the limitations of lexicons. Let us mention Linguistica (Goldsmith 2001), Morfessor (Creutz and Lagus 2005), or, more recently, Cotterell and Schütze (2017)'s models. These systems are applicable to any language, however they are more effective for languages with concatenative morphology such as English, German and French. Morphological analysis can also be carried out by symbolic parsers, most of them developed by linguists; for a panorama, see (Bernhard et al. 2011).

Lexical resources with derivational annotations can replace or supplement morphological parsers in the NLP pipeline if their lexical coverage is large enough and if their features are sufficiently rich and varied. In its meticulous and exhaustive report, Kyjánek (2018) produces a typology describing the structure and coverage of 30 recent derivational resources for Romance (including Latin), Germanic and Slavic languages. The reader should refer to this work to get a clear idea of the existing derivational databases (DDBs) and lexicons with derivational annotations.

The lack of large-scale derivational resources of French motivated the development, from 2011, of a prototype database Démonette$_{v1}$ (Hathout and Namer 2014a, 2016). Démonette$_{v1}$ describes derivational families made up of verbs, agent and action nouns and modality adjectives. Three objectives were pursued: (1) use DériF's analyses (Namer 2009, 2013) to produce a resource whose inputs are derivational relations between two words W_1 and W_2, labelled with linguistically grounded fetures, including semantic annotations; (2) complete these $W_1 \rightarrow W_2$ derivations by relations between derivational family members provided by the analogic model implemented in Morphonette (Hathout 2009); (3) define an extensible and redundant architecture, which can be fed by varied and heterogeneous morphological resources. The design of the Démonette$_{v2}$ database (§.3) is based on the experience gained during the development of Démonette$_{v1}$. The aim is to produce a lexicon whose descriptions (morphological, phonological, frequency, and especially semantic) will be useful for NLP, but will also serve as a reference for several audiences (research in morphology, university teaching, academic or speech therapy practice, just to cite a

few). The structure of the database must allow (semi-)automatic acquisition from existing resources, and must be robust enough to be able to include any new type of derivation. We therefore need an architecture based on theoretical principles that ensure a uniform representation of regular derivation (words where meaning and form deduce from each other) and non-canonic derivation, which infringe form-meaning compositionality. For this purpose, Démonette$_{v2}$ applies the theoretical principles borrowed from lexeme- and paradigm-based approaches to word formation (WF), summarized in §.2.

2 Démonette's theoretical background

Two major facts have independently contributed to recent evolution in WF, and have therefore influenced the content and organization of derivational resources: (1) the adoption of the lexeme as a unit, and (2) the structuration of the morphological lexicon into paradigms.

2.1 Morphemes, and form-meaning non-compositionality

Morpheme-based morphological traditions, whether concatenative (Item and Arrangement) or functional (Item and Process) (Hockett 1954), have long been taken as models for the development of automatic derivation tools. However, the limits of morpheme-based morphology have been widely discussed in the literature (Aronoff 1976, Anderson 1992, Fradin 2003): the most significant drawback concerns the rigidity of the morpheme, a unique and minimal combination of form and meaning which cannot easily adapt to non canonical derivation (Corbett 2010). In these frameworks, the analysis of words whose meaning and form do not coincide becomes (very) complex. One example is *zero affixation* or *conversion* (Tab.1-a) (Tribout 2012), characterized as "formal undermarking" of the derivative with respect to its base by Hathout and Namer (2014b) (the derived form is identical to the base form but its semantic content is more complex). On the other hand, *parasynthetic* derivatives (Tab.1-b,c) (Hathout and Namer 2018), are said to be "over-marked" because one of their formal parts does not play a role in the construction of their meaning. Finally, the derivational relations obtained by *affix replacement* (Booij and Masini 2015) are both "under- and over-marked" with respect to each other: in Tab.1-d, Lex$_2$ is constructed by replacing *-ism* in Lex$_1$ by *-ist* (and vice versa). Non-canonical derivations also include processes that regularly produce two series of words with the same shape but different meanings, or with distinct forms but the same meaning. In the first case, (absence of formal markdown) the derivative is *polysemic* (in French, cf. Tab.1-e, *-eur* suffixed nouns denote either humans or artifacts). The second case corresponds to *morphological variation* or *competition*. Here, the absence of semantic markdown corresponds to what Thornton (2012) calls *overabundance*: for instance, in Italian, Tab.1-f, prefixes *s-* and *de-* compete to form adjective-based verbs, cf. (Todaro 2017).

	formation	lgge	Lex$_1$	Lex$_2$
a	conversion	eng	*nurse*$_N$	*nurse*$_V$
b	parasynthesis	fra	*banque*$_N$ 'bank'	*interbancaire*$_A$ 'between banks'
c			*département*$_N$ 'department'	*interdépartemental*$_A$ 'between departments'
d	affix replacement	eng	*altruism*$_N$	*altruist*$_N$
e	polysemy	fra	porter$_V$ 'carry'	*porteur*$_{Nm,[hum]OR[artif]}$ 'carrier'
f	overabundance	ita	*compatto*$_A$ 'compact'	*scompattare*$_V$ or *decompattare*$_V$ 'uncompact'

Table 1: Different types of meaning-form discrepancies in Lex$_1$/Lex$_2$ derivational relations.

2.2 Lexemes, and non-binary or non-oriented rules

Abandoning the morpheme in favour of the *lexeme* solves some problems that arise from meaning non-compositionality. Unlike the morpheme, lexeme is not a concrete minimal unit. It is actually an abstract object (an uninflected word, in the simplest cases) that records the common properties of the inflectional

paradigm it stands for, in the form of an autonomous three-dimensional structure: (1) a phonological form (the stem); (2) a part-of-speech; (3) a meaning. Unlike morpheme concatenation rules which apply an affixal function to a morphological structure, *word formation rules* (WFRs) are oriented relations between two lexemes or schemas, as in (1). WFRs apply independently and simultaneously to all three levels of description allowing the formal exponent to vary for a same semantic type of derivative, and vice versa.

In particular, this evolution solves part of the problems illustrated in Tab.1. For conversion (Tab.1-a), as shown in (1), the rule only modifies the semantic content and the part-of-speech, leaving the formal values of the related lexemes unchanged; as for polysemy (Tab.1-e), nothing prevents two distinct rules to derive word-types with different semantic content using the same formal exponent; as for overabundance (Tab.1-f), two different WFRs can produce different formal realizations for the same semantic value.

(1) $\begin{bmatrix} /nɜːs/ \\ N \\ \text{'nurse'} \end{bmatrix} \rightarrow \begin{bmatrix} /nɜːs/ \\ V \\ \text{'ACTING as a nurse'} \end{bmatrix}$

However WFRs are designed to connect a derivative to its base. They are not designed to describe indirect relations, such as (Tab.1-d). For the same reason, lexeme-based models are not able to describe parasynthetic derivation (Tab.1-b,c) where, for a given prefixation process (e.g. *inter-*), the suffix exponent is not unique (*-aire* in *banque* → *interbancaire*, but *-al* in *département* → *interdépartemental*), and the suffix value cannot be determined by neither the form or the meaning of the base.

2.3 Paradigms, and partially motivated relations

Derivational paradigms overcomes the limitations of the lexeme-based morphology where derivational relations are restricted to binary and oriented base$_W$ → derived$_W$ connections (for a panorama, see Štekauer (2014)). In a paradigmatic framework (Bonami and Strnadová 2019), the central unit is the *derivational family*, i.e. a structured set of lexemes[1], whose form and meaning depend on each other: all the members of a family are interconnected. Two families belong to the same paradigm when they line up; in this alignment, members of the same rank or position maintain in their respective families the same form and meaning relations with the other members of their family, and are therefore part of the same *derivational series* (Hathout 2011). Families may align partially. In such a framework, directly and indirectly related word pairs are both described in the same way by means of non-oriented schemata as in (2). This schema decribes the relation between *altruist* and *altruism* of Tab.1-d, where X is set for their common subsequence /æltrʊ/. Semantically, the mutual motivation of the two nouns is described by means of the "@1" and "@2" indexes: *altruism* is the "IDEOLOGY DEFENDED by (an) altruist", which in turn is a "FOLLOWER of altruism".

(2) $\begin{bmatrix} /Xɪst/ \\ N \\ @1{:}\text{'FOLLOWER of @2'} \end{bmatrix} \leftrightarrow \begin{bmatrix} /Xɪzm/ \\ N \\ @2{:}\ \text{'IDEOLOGY DEFENDED by @1'} \end{bmatrix}$

(*altruist*, *altruism*) is a partial family that belongs to a sub-paradigm of the paradigm resulting from the stacking of triplets like the ones presented in Tab.2. Each triplet connects a (proper) noun denoting an entity (X), a noun of ideology (Xism) valuing that entity, and a human noun (Xist) denoting a person supporting that ideology. In his *Cumulative Patterns* Bochner (1993) represents these paradigmatic relations in the form of ternary schemata as in (3)[2].

(3) $\left\{ \begin{bmatrix} /Xɪst/ \\ N \\ @1{:}\text{'FOLLOWER of @2,} \\ \text{ENDORSING @3'} \end{bmatrix}, \begin{bmatrix} /Xɪzm/ \\ N \\ @2{:}\ \text{'IDEOLOGY} \\ \text{DEFENDED by @1,} \\ \text{PROMOTING @3'} \end{bmatrix}, \begin{bmatrix} /X/ \\ \text{PrN or N} \\ @3{:}\ \text{'ENTITY} \\ \text{PROMOTED by @2,} \\ \text{ENDORSED by @1'} \end{bmatrix} \right\}$

[1]The notion of paradigm does not necessarily imply that of lexeme. Nevertheless, we are only interested here in this type of unit.

[2]Various other theoretical approaches have been proposed to represent paradigms in derivation by Koenig (1999), Booij (2010), Spencer (2013), Antoniova and Štekauer (2015) to only cite a few.

X: Valued Entity	Xist: Follower	Xism: Ideology
Calvin	calvinist	calvinism
race	racist	racism

Table 2: (X, Xist, Xism) paradigm in English.

However, some questions raised by the derivations in Tab.1-(b,c) remain unanswered. One of them is the variable value of the suffix on the adjective prefixed by *inter-*: *interbanc**aire***, *interdépartement**al***, but also *interocéan**ique*** 'between oceans' or *intercoral**lien*** 'between corals'. Moreover, there is a meaning-form asymetry because the suffix does not contribute to the adjectival meaning, basically, a spatial interval between two or more concrete entities ('between several X') where X is, respectively, *banque*, *département*, *océan* and *corail*. When observing the derivational family of these adjectives (Tab.3), we can see that the suffix that shows up in the prefixed adjective is the same as the one of the relational adjective ('of X') of all these nouns.

In a way, the adjective in *inter-* has two bases: the noun X is its semantic base, and the adjective Xsuf its formal base. In other words, the construction of *inter*Xsuf$_A$ requires simultaneous access to the semantic properties of X$_N$, and the formal properties of Xsuf$_A$.

X$_N$	Xsuf$_A$: 'of X'	*inter*Xsuf$_A$: 'between several Xs'
banque	*bancaire*	*interbancaire*
département	*départemental*	*interdépartemental*
océan	*océanique*	*interocéanique*
corail	*corallien*	*intercorallien*

Table 3: (X, Xsuf, *inter*Xsuf) paradigm in French.

An access to the derivational family of the prefixed adjective is therefore necessary for the description and prediction of its properties. However, "classical" paradigmatic organizations such as the ones we have just presented are too rigid to express the double ascendancy of the *inter*Xsuf adjectives. Classical paradigmatic systems are actually designed to describe regularities that hold at all three levels: formal, categorial and semantic. These paradigms are therefore unable to capture the regularities that involve lexemes with a mismatch between form and meaning, like the ternary relations in Tab.3. To properly describe and predict this type of discrepancy, the semantic and formal relations must be described and accessed separately, as they do in ParaDis.

2.4 ParaDis

As shown in the previous section, the principles of lexeme-based and paradigmatic approaches to derivation are both required in order to provide WF models and resources with sufficient descriptive and predictive power. However, they remain unable to account for asymetrical formations as in Tab.3-b,c. Far from being exceptional, such formations occur in a large part of the prefixed denominal adjectives of French (and other European languages): they describe a spatial relation (*inter-*, *intra-*, *sous-*, *sur-*, ...), adversativity (*anti-*), quantification (*mono-*, *bi-*, *pluri-*,...), etc. Other types of derived words display comparable over-marks with respect to their bases: for example, in French, verbs like *scolariser*$_v$ 'get into school' are formally formed by suffixation in *-iser* on an adjectival base (*scolaire*$_A$ 'of school') while their semantic content is built on the meaning of the base noun of this adjective (*école*$_N$ 'school').

We therefore need a model that grasps the paradigmatic regularities blurred by the many form-meaning discrepancies, by transposing the main contribution of lexeme-based morphology (independent formal, categorial and semantic levels do representations) to the paradigmatic organization of the lexicon (access to all the members of a derivational family). In other words, the model must combine a morpho-phonological paradigmatic network (in order for example to predict the formal motivation of *inter*Xsuf

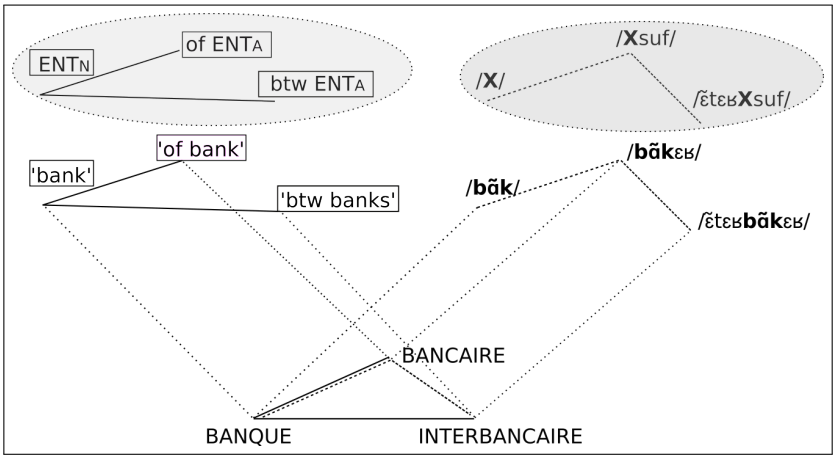

Figure 1: ParaDis: Representation of the (X, Xsuf, *inter*Xsuf) unbalanced paradigm.

with respect to Xsuf) and a morpho-semantic paradigmatic network (able for instance to predict the semantic motivation of *inter*Xsuf with respect to X) in order to properly describe and predict these adjectives.

This is precisely what we propose in ParaDis "Paradigms vs Discrepancies" (Hathout and Namer 2018). The model is based on the assumption that a derivational paradigm behaves as a kind of generalization of the lexeme's ternary structure: it contains the same three-level organization. The premise is that, if morphological regularities are paradigmatic, then the morpho-semantic, morpho-categorial and morpho-formal levels in correspondence with these paradigms are themselves paradigms. In other words, ParaDis brings to the semantic, categorial and formal levels the organizational principles of classical paradigm-based WF models. This system therefore includes a (morpho-)formal paradigm, a (morpho-)categorial paradigm and a (morpho-)semantic paradigm, whose junction is the morphological paradigm they are in correspondence with. This morphological paradigm is the abstract combination of the other three components, just as the lexeme is the abstract combination of a formal, categorial and semantic descriptions.

The independence of the formal, categorial and semantic paradigms allows a three-dimensional description of asymetric derivations like *interbancaire*, cf. Fig.1. For sake of readability, we have merged the categorial and the semantic levels. The formal paradigm (gray oval on the right) is an alignment of families of forms; families are represented as connected graphs, where each edge expresses a formal motivation between two phonological sequences. The semantic paradigm (gray oval on the left) is an alignment of families of concepts; families are represented as connected graphs, where each edge expresses a semantic motivation between two semantic values. These graphs are incomplete and they differ from each other. In the semantic paradigm, the semantic values that represent the spatial interval ('btw ENTITIES') and the relation ('of ENTITY') are not deductible from each other, and therefore they are not related. Likewise, in the formal paradigm, the two unrelated formal patterns /X/ and /ɛ̃tɛʁXsuf/ are not interpredictable. In the morphological paradigm (bottom), the relation between *banque* and *bancaire* is regular (displayed by a double line): it inherits a semantic motivation from the semantic paradigm, and a formal motivation from the formal paradigm. Conversely, there is only a formal relation between *bancaire* and *interbancaire* (displayed by hyphens), and only a semantic relation (solid line) between *banque* and *interbancaire*. The other families of Tab.3 are analyzed in the same way.

In the next section, we show how Démonette$_{v2}$'s implements the main features of ParaDis.

3 The Démonette$_{v2}$ derivational database

The organization of Démonette$_{v2}$ is original: an entry in the DDB corresponds to a *derivational relation* between two lexemes belonging to the same family, but not necessarily in a base/derivative relationship. The DDB is thus based on the theoretical assumptions summarized in § 2 which consider that the lexeme is the fundamental morphological unit and that the derivational construction fulfills two functions: (1)

create new lexemes and (2) establish semantic and formal relations of motivation between the lexemes present in the lexicon.

In addition to the initial contribution of the 96,000 entries of Démonette$_{v1}$ (Hathout and Namer 2014a, Namer et al. 2017), the content of Démonette$_{v2}$ is obtained by migrating existing derivational resources, developed and validated by morphologists. These resources where selected because of their availability, complementarity and richness of description (morphological annotations and, for most of them, semantic and phonological features). Their processing is scaled according to the complexity of their migration in the format of Démonette$_{v2}$. These resources amount to 183,000 entries, most often in the form of annotated (base$_W$, derived$_W$) word pairs, corresponding to ca. 120 derivational processes by conversion, suffixation (*-ard*, *-ariat*, *-at*, *-âtre*, *-el*, *-aie*, *-iser*, *-erie*, *-esque*, *-esse*, *-eur*, *-eux*, *-iste*, ...), or prefixation (*a-*, *anti-*, *bi-*, *co-*, *contre-*, *dé-*, *é-*, *extra-*, *hyper-*, *hypo-*, *in-*, *infra-*, *inter-*, ...). The migration often involves a reanalysis of the original base/derivative connections in order to produce a description compatible with Démonette$_{v2}$'s principles. Moreover, new information, new connections and new lexemes may be added (semi-)automatically in order to extend derivational families.

3.1 Overview

Démonette$_{v2}$ implements the fundamental features of ParaDis. In other words, the structure of this database is based on the following principles, some of which being already implemented in the Démonette$_{v1}$ prototype.

- each *entry* describes a relation between two lexemes of a derivational family: the same lexeme therefore intervenes in as many entries of the base as it has relations within its family,
- each entry is *annotated* with respect to the relation and to each of the two related lexemes,
- the description of a *lexeme* is stable because it is independent of the connections it takes part. It consists of a standardized written form, a part-of-speech, an inflectional paradigm (in IPA format), and an ontological type, selected among the 25 WordNet *Unique Beginners (UB)* (Miller et al. 1990)),
- *relations* are defined by three independent sets of properties: structural ones (characterization of the morphological connection itself), formal ones (formal pattern of each lexeme and stem variation, if any) and semantic ones (semantic type of the relation and glosses that mutually defines the two lexemes).

The remaining of the paper presents the architecture of Démonette$_{v2}$ and its formal, structural and semantic parts. The reader can refer to (Namer et al. 2017) for a presentation of the morpho-phonological properties. We mainly show how this structure allows families to be grouped into formal, semantic and derivational networks and will ultimately provides a large-scale description of the paradigmatic organization of the morphologically complex lexicon that takes into account meaning-form discrepancies.

3.2 Regular Paradigms in Démonette$_{v2}$

Let us consider the five families of Tab.4. Each one is built around a verb predicate (*laver* 'wash'), and includes an iterative verb (*relaver* 're-wash'), the action nouns of the two predicates (*lavage* 'washing', *relavage* 're-washing'), and an adjective indicating potentiality (*lavable* 'wash-able'). In French, action nouns may be constructed by conversion rule (*découper* / *découpe*) or suffixation, in which case several exponents are available (*-age*, *-ment*, *-ion*, *-ure*, ...). However, the same formal process is used for the nominalization of the simple predicate and the iterative predicate. The derivational relations between the five members of each family in Tab.4 are all regular because they all are formally and semantically motivated. These relations form complete oriented graphs with 2×10 edges. Each edge is an entry in the Démonette$_{v2}$ DDB.

X_V	X(suf)$_N$	*re*X$_V$	*re*X(suf)$_N$	X*able*$_A$
laver ‘wash’	lav**age**	relaver	relav**age**	lavable
classer ‘rank’	classe**ment**	reclasser	reclasse**ment**	classable
planter ‘plant’	plantat**ion**	replanter	replantat**ion**	plantable
souder ‘weld’	soud**ure**	resouder	resoud**ure**	soudable
découper ‘cut (out)’	découpe	redécouper	redécoupe	découpable

Table 4: (X_V, *re*X_V, Xsuf, *re*Xsuf, X*able*) families in French.

Tab.5 describes the way each relation in the family of *laver* is labelled in Démonette$_{v2}$[3]. This description involves four features: `Ori(entation)` and `Co(mplexity)` identify the relation's structure, whereas `Sch(ema)`$_{L1}$ and `Sch(ema)`$_{L2}$ encode the formal patterns L1 and L2 match within this relation. For a given (L1, L2) entry, `Ori` indicates whether L1 is the ancestor of L2 (`a2d` value), whether L2 is the ancestor of L1 (`d2a` value) or whether there is an `ind(irect)` relation between them. Note that the feature `Ori=ind` characterizes formations with an affix replacement (Tab.1-d), for example in the follower/ideology relations as in Tab.2. The `Co` feature describes the number of morphological steps necessary to reach L2 from L1. In the case of a regular derivation, its value is `si(mple)` when one of the two lexemes is the base of the other, or when both have a common base; the value `co(mplex)` is used in the other cases. `Sch`$_{L1}$ and `Sch`$_{L2}$ indicate which exponents are needed in the relation to go from L1 to L2: X represents the sequence they have in common in this context.

L1	L2	Sch$_{L1}$	Sch$_{L2}$	Ori	Co	L1	L2	Sch$_{L1}$	Sch$_{L2}$	Ori	Co
laver	*lavage*	X	Xage	a2d	si	*laver*	*relavage*	X	reXage	a2d	co
laver	*relaver*	X	reX	a2d	si	*laver*	*lavable*	X	Xable	a2d	si
lavage	*relavage*	X	reX	a2d	si	*lavage*	*relaver*	Xage	reX	ind	si
lavage	*lavable*	Xage	Xable	ind	si	*relavage*	*relaver*	Xage	X	d2a	si
relavage	*lavable*	reXage	Xable	ind	co	*relaver*	*lavable*	reX	Xable	ind	si

Table 5: Démonette$_{v2}$ – Encoding structural and formal properties in the family of *laver*

When `Complexity=simple`, the formal description of the relation is coupled with a semantic annotation (Tab.6). This provides information on the semantic value of the relation (`RSem`), for instance `syn(onymy), iter(ation)` or `pot(entiality)`, for the relations in Tab.4. Semantic descriptions also include a paraphrase defining the two related words with respect to each other. For instance, the gloss for (*lavage*, *lavable*) is: “`One can perform` lavage `on something if it is` lavable”). The generalization of such paraphrases (col. 6) is obtained by replacing the words L1 and L2 by their ontological types (cols. 3 and 4). The derivational relations in the other families of Tab.4 are annotated structurally, formally and semantically in the same way. The generalization made on all the features allows families to align and paradigmatic regularities to emerge.

3.3 Meaning-form discrepancies in Démonette$_{v2}$

So far, we have shown how the architecture of Démonette$_{v2}$ allows for the representation of classical and regular derivations (*laver/lavage*), but also derivations with some meaning-form mismatches summarized in the Tab.1: conversion and overabundance are dealt with by the autonomy of features in the (L1, L2) relation, the semantic types (`RelSem`) being independent from the formal structures (`Sch`$_{Li}$). It is also able to deal with polysemy and affix replacement, , the latter being identified by the feature `Orientation=indirect`.

[3]For reasons of space, word pair are listed in only one direction. The description of $L2 \rightarrow L1$ is symmetrical to that of $L1 \rightarrow L2$: the values of Sch$_{L1}$ and Sch$_{L2}$ and of TySem$_{L1}$ and TySem$_{L2}$ are inverted (Tab.5, Tab.6); the value `a2d` substitutes for `d2a` and vice-versa for the feature `Complexity` (Tab.5); the values of the other features are unchanged.

L1	L2	TySem$_{L1}$	TySem$_{L2}$	RSem	Def_abs
laver	*lavage*	Act$_{V1}$	Act$_{N2}$	syn	'To Act$_{V1}$ sth is to perform Act$_{N2}$'
relavage	*relaver*	Act$_{V1}$	Act$_{N2}$		'To Act$_{V2}$ sth is to perform Act$_{N1}$'
laver	*relaver*	Act$_{V1}$	Act$_{V2}$	iter	'To Act$_{V1}$ smth several times is to Act$_{V2}$ it'
lavage	*relavage*	Act$_{N1}$	Act$_{N2}$		'To perform several Act$_{N1}$ is to perform Act$_{N2}$'
lavage	*relaver*	Act$_{N1}$	Act$_{V2}$		'To perform several Act$_{N1}$ is to Act$_{V2}$'
laver	*lavable*	Act$_{V1}$	Mod$_{A2}$	pot	'One can Act$_{V1}$ sth if it is Mod$_{A2}$'
lavage	*lavable*	Act$_{N1}$	Mod$_{A2}$		'One can perform Act$_{N1}$ on sth if it is Mod$_{A2}$'
relaver	*lavable*	Act$_{V1}$	Mod$_{A2}$		'One can Act$_{V1}$ several times sth if it is Mod$_{A2}$'

Table 6: Démonette$_{v2}$ – Encoding semantic properties in the family of *laver*

In Démonette$_{v2}$, derivational families can be reconstructed from the network of direct and indirect relations that connect its members. Then, families can be grouped into semantic (resp. formal) paradigms if the (L1, L2) relations included in the families are aligned (`Co` and `Ori` values are identical), and belong to the same semantic (resp. formal) series, i.e. share the same values for TySem$_{L1}$, TySem$_{L2}$, `RSem` and `Def_abs` (resp. for Sch$_{L1}$ and Sch$_{L2}$). The paradigms may include sub-paradigms made up of partial families.

Démonette$_{v2}$ can also represent paradigms with heterogeneous connections, such as those in Tab.3. The set of features we use allows for the compartmentalization of the descriptions into formal, structural, semantic and phonological levels. The analysis of meaning-form discrepancies then does not require any modification in the architecture. We only need two additional values, `f(ormal)-m(otivation)` and `s(emantic)-m(otivation)`, for the attribute `Complexity`. We illustrate their role with Tab.7. Each column corresponds to an entry of Démonette$_{v2}$. These entries connect the members of the morphological family of *banque*$_N$, as displayed in the lower part of Fig. 1. When the (L1, L2) relation is only formally motivated, it is encoded with the `f-m` value (col. 4) and does not involve a semantic description. On the other hand, the value `s-m` (col. 3) signals a semantically grounded relation with no formal motivation. Recall that regular relations like *banque/bancaire* are noted `Co=simple` (col.1): in other words, this value merges `f-m` and `s-m` [4].

L1 – L2	*banque – bancaire*	*banque – interbancaire*	*bancaire – interbancaire*
Sch$_1$/Sch$_2$	X$_N$/Xaire$_A$	X$_N$/interXaire$_A$	X$_A$/interX$_A$
Ori	a2d	a2d	a2d
Co	si	**s-m**	**f-m**
SemRel	relation	space interval	–
Def.	'Smth bancaire pertains to the bank'	'Smth interbancaire relates to several banks'	–

Table 7: Démonette$_{v2}$ entries for the family of *banque*$_N$

With `s-m` and `f-m`, Démonette$_{v2}$ can independently represent formal and semantic paradigms just as in ParaDis and thus becomes a large-scale formalization of this model: a relation with `Complexity=f-m` only belongs to the formal network (no semantic counterpart) while a relation with `Complexity =s-m` only belongs to the semantic network.

4 Conclusion

We have presented Démonette$_{v2}$ and its theoretical background. This resource is under development, and therefore the results we have presented are still partial. The WF principles we choose follow from the

[4]The same features are used for the description of the members of the other families in Tab.3.

objectives of the database. Our goal is to provide a semantically and formally homogeneous description of morphologically constructed French words, formed by regular derivations as well as non-canonical WFR. One way to achieve this goal is to combine the contributions of lexeme-based morphology and paradigmatic models of derivation.

This work also shows how a lexical resource and a theoretical model can cross-fertilize even if our presentation mainly focused on the theoretical foundations of Démonette$_{v2}$. We therefore omitted other aspects of the database: among them, the edition and visualization platform, the devices operating a (partial) automatisation for the semantic annotations of the lexemes and the relations, the elaboration of the glosses, the automatic extension of families. For the last task, several approaches are envisaged, including the formalization of linguistic reasoning, the implementation of neural networks or the application of formal concept analysis (Leeuwenberg et al. 2015). We have also left aside the conversion of the content of Démonette$_{v2}$ in order to meet the needs of the different uses of the database, by reseachers and students interested in morphology, elementary school teachers, speech-language pathologists specialized in language acquisition disorders. All these topics are ongoing research. Their results will be published in future.

Acknowledgments

This work benefited from the support of the project DEMONEXT ANR-17-CE23-0005 of the French National Research Agency (ANR). We wish to thank the partners of DEMONEXT, and especially Lucie Barque and Pauline Haas who have also taken part in the results presented in this paper.

References

Stephen R. Anderson. 1992. *A-Morphous Morphology*. Cambridge University Press, Cambridge, UK.

Vesna Antoniova and Pavol Štekauer. 2015. Derivational paradigms within selected conceptual fields – contrastive research. *Facta Universitatis, Series: Linguistics and Literature* 13(2):61–75.

Mark Aronoff. 1976. *Word Formation in Generative Grammar*. Linguistic Inquiry Monographs. MIT Press, Cambridge, MA.

Delphine Bernhard, Bruno Cartoni, and Delphine Tribout. 2011. A task-based evaluation of French morphological resources and tools. *Linguistic Issues in Language Technology* 5(2).

Harry Bochner. 1993. *Simplicity in generative morphology*. Mouton de Gruyter, Berlin & New-York.

Olivier Bonami and Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology* 29(2):167–197.

Geert Booij. 2010. *Construction Morphology*. Oxford University Press, Oxford.

Geert Booij and Francesca Masini. 2015. The role of second order schemas in the construction of complex words. In Laurie Bauer, Lívia Körtvélyessy, and Pavol Štekauer, editors, *Semantics of complex words*, Springer, Heidelberg, volume 47, pages 47–66.

Greville G. Corbett. 2010. Canonical derivational morphology. *Word Structure* 3(2):141–155.

Ryan Cotterell and Hinrich Schütze. 2017. Joint semantic synthesis and morphological analysis of the derived word. *CoRR* abs/1701.00946. http://arxiv.org/abs/1701.00946.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical Report A81, Helsinki University of Technology.

Bernard Fradin. 2003. *Nouvelles approches en morphologie*. PUF, Paris.

John Goldsmith. 2001. Unsupervised learning of the morphology of natural language. *Computational Linguistics* 27(2):153–198.

Nabil Hathout. 2009. Acquisition of morphological families and derivational series from a machine readable dictionary. In Fabio Montermini, Gilles Boyé, and Jesse Tseng, editors, *Selected Proceedings of the 6th Décembrettes: Morphology in Bordeaux*. Cascadilla Proceedings Project, Somerville, MA.

Nabil Hathout. 2011. Une approche topologique de la construction des mots : propositions théoriques et application à la préfixation en *anti-*. In *Des unités morphologiques au lexique*, Hermès Science-Lavoisier, Paris, pages 251–318.

Nabil Hathout and Fiammetta Namer. 2014a. Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology* 11(5):125–168.

Nabil Hathout and Fiammetta Namer. 2014b. Discrepancy between form and meaning in word formation: the case of over- and under-marking in French. In Franz Rainer, Wolfgang U. Dressler, Francesco Gardani, and Hans Christian Luschützky, editors, *Morphology and meaning*, John Benjamins, Amsterdam, pages 177–190.

Nabil Hathout and Fiammetta Namer. 2016. Giving lexical resources a second life: Démonette, a multi-sourced morpho-semantic network for french. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Nabil Hathout and Fiammetta Namer. 2018. La parasynthèse à travers les modèles : des RCL au ParaDis. In Olivier Bonami, Gilles Boyé, Georgette Dal, Hélène Giraudo, and Fiammetta Namer, editors, *The lexeme in descriptive and theroretical morphology*, Language science Press, Berlin, Empirically Oriented Theoretical Morphology and Syntax, pages 365–399. http://langsci-press.org/catalog/book/165.

Charles Francis Hockett. 1954. Two models of linguistc descriptions. *Words* 10:210–234.

Jean-Pierre Koenig. 1999. *Lexical Relations*. CSLI Publications, Stanford, CA.

Lukáš Kyjánek. 2018. Morphological resources of derivational word-formation relations. Technical Report 61, ÚFAL - Charles University, Prague.

Artuur Leeuwenberg, Aleksey Buzmakov, Yannick Toussaint, and Amedeo Napoli. 2015. Exploring pattern structures of syntactic trees for relation extraction. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)* 9113:153–168. https://doi.org/10.1007/978-3-319-19545-2-10.

Georges A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography* 3(4):335–391.

Fiammetta Namer. 2009. *Morphologie, lexique et traitement automatique des langues : L'analyseur DériF*. Hermès Science-Lavoisier, Paris.

Fiammetta Namer. 2013. A rule-based morphosemantic analyzer for French for a fine-grained semantic annotation of texts. In Cerstin Mahlow and Michael Piotrowski, editors, *SFCM 2013*, Springer, Heidelberg, CCIS 380, pages 93–115.

Fiammetta Namer, Nabil Hathout, and Stéphanie Lignon. 2017. Adding morpho-phonology into a french morpho-semantic resource: Demonette. In Eleonora Litta and Marco Passarotti, editors, *Proceedings of the First Workshop in Resources and Tools for Derivational Morphology (DeriMo),*. EDUCatt, Milano, Italy, pages 49–60.

Andrew Spencer. 2013. *Lexical relatedness*. Oxford University Press, Oxford.

Pavol Štekauer. 2014. Derivational paradigms. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Derivational Morphology*, Oxford, Oxford University Press, Oxford, pages 354–369.

Anna M Thornton. 2012. Reduction and maintenance of overabundance. a case study on italian verb paradigms. *Word Structure* 5(2):183–207.

Giuseppina Todaro. 2017. *Nomi (e aggettivi) che diventano verbi tramite prefissazione: quel che resta della parasintesi*. Ph.D. thesis, Tesi di dottorato, Università Roma Tre et Université Toulouse Jean-Jaurès.

Delphine Tribout. 2012. Verbal stem space and verb to noun conversion in french. *Word Structure* 5(1):109–128.