A scenario-based approach to predictability in derivation

Daniele Sanacore
CLLE, Université de Toulouse

Nabil Hathout
CLLE, Université de Toulouse

Fiammetta Namer
ATILF, Université de Lorraine

Abstract This article explores the notion of *derivational paradigm*. Although several studies have proposed a paradigmatic approach to derivational morphology, we don't know yet what derivational paradigms look like. A key feature of paradigms in inflection is the mutual predictability of the paradigm cells in terms of the content that they express, but we don't know yet how predictability works in the derivational lexicon. In order to explore predictability in derivation, we propose to use *scenarios* (i.e. prototypical representations of real-world situations). The idea is to build scenarios using short stories produced by Large Language Models (LLMs). We create stories containing pairs of lexemes belonging to the same word family; the regular content of the stories and the participants that frequently co-occur in them determine the prototypical participants of the scenarios. The participants of the same scenario can be considered as semantically interpredictable and may be realized by lexemes belonging to the same derivational paradigm.

Keywords: morphology; derivational paradigms; paradigmatic families; frame semantics; LLMs

1 Introduction

This article explores the notion of derivational paradigm in French. The idea of extending the notion of paradigm, traditionally associated with inflection, to derivation has been discussed in several articles (Bauer 2019; Pounder 2000; Štekauer 2014; Boyé & Schalchli 2016; Hathout & Namer 2019, among others) and has been the subject of numerous workshops and papers. However, it seems that a consensus on the definition of *derivational paradigm* has not yet been reached among morphologists.

A central aspect for the structure of inflectional paradigms is the interpredictability of the elements that they contain in terms of morphosyntactic content (Bonami 2014; Beniamine 2018). On the other hand, predictability in derivation has not been extensively explored yet (McNally et al. 2024). These observations are the starting point of our paper. We address two main questions: (i) what do derivational paradigms look like? (ii) how can predictability be investigated in the derivational lexicon?

To explore predictability in derivation, we take semantics as starting point and we consider that recurrent and predictable semantic relations are central to the structure of derivational paradigms (Štekauer 2014; Bonami & Strnadová 2019). We propose to explore the predictability of semantic relations by taking inspiration from the notions of *scenario* (Sanford & Garrod 1998; Erk & Herbelot 2024) and *frame*(Fillmore 1976; Petruck 1996). Scenarios, like frames, are conceptual structures that represent real world situations involving a set of prototypical participants. For example, a commercial transaction scenario generally involves (at least) a buyer, a seller, some goods that are sold by

the seller and some money in exchange. We consider that the way frames and scenarios work, with several elements closely related that function as a whole, is quite close to our idea of derivational paradigms as being structured by a bundle of predictable semantic relations that can be realized morphologically (Sanacore 2023).

In the theoretical framework of frame semantics, it is assumed that frames are instantiated in language use: for example, the sentence *Abby bought a car from Robin for 5000 dollars* instantiates a commercial transaction frame (Petruck 1996: p.1). The same frame can be instantiated by a large set of corpus sentences (i.e. all sentences that "evoke" a commercial scenario in the mind of the reader). In this article, we propose to estimate the predictability of semantic relations in derivation by using stories generated by Large Language Models (LLMs). We consider that the regular participants in the stories produced starting from pairs of derivationnally related words *word1* and *word2* correspond (a) to the concepts that are semantically predictable given the relation between these two words and (b) to the potential cells in the derivational paradigm that features *word1* and *word2*. In other words, the regular participants of the stories correspond to the prototypical participants of the scenario where *word1* and *word2* are inscribed. In our proposal, derivational paradigms are thus delimited by scenarios that contain participants and activities that are highly predictable one from the other in stories. We determine interpredictability in stories using conditional probability. Firstly, we compute the probability of a prototypical participant B (e.g. buyer) to appear in a story built on a pair of words A (e.g. HERBORISTE-HERBORISTERIE). Secondly, we compute the probability of a prototypical participant C (e.g. shop) to appear in a story given the presence of a participant B (e.g. buyer).

2 Derivational predictability

According to (Fradin 2020: among others), the main difference between inflectional and derivational paradigms is the nature of their content: inflectional paradigms are determined by morphosyntactic features relevant to the grammar (e.g., the form *laverai* of the verb LAVER 'wash' realizes the features FUT.1SG), while derivational paradigms realize conceptual categories that are relevant for the lexicon such as agent (e.g., LAVEUR 'washer') or instrument (e.g., ASPIRATEUR 'vacuum cleaner').

A fundamental property of inflectional paradigms is the interpredictability of the elements that they contain in terms of morphosynctactic content (Bonami 2014; Beniamine 2018). For example, in the paradigm of the verb LAVER, the presence of the form *lave*, which realizes PRES.1SG, and of the form *laverai*, which realizes FUT.1SG, predict each other (Hathout & Namer 2022: 156). Interpredictability is central for the implicative structure of inflectional paradigms. On the other hand, to the best of our knowledge, predictability has not been investigated for derivation and our work aims to bring a contribution to this question.

2.1 Predictability in derivational families

Several authors (Dokulil 1982; Stump 1991; Gaeta 2022, among others) use the term *derivational paradigm* to refer to word families composed of lexemes that share a common semantic core and are connected by direct or indirect derivational relations. We will call them *derivational families*, following Bonami & Strnadová (2019).

If derivational paradigms correspond to full word families, several questions arise concerning the interpredictability of their members. For example, consider the family of the

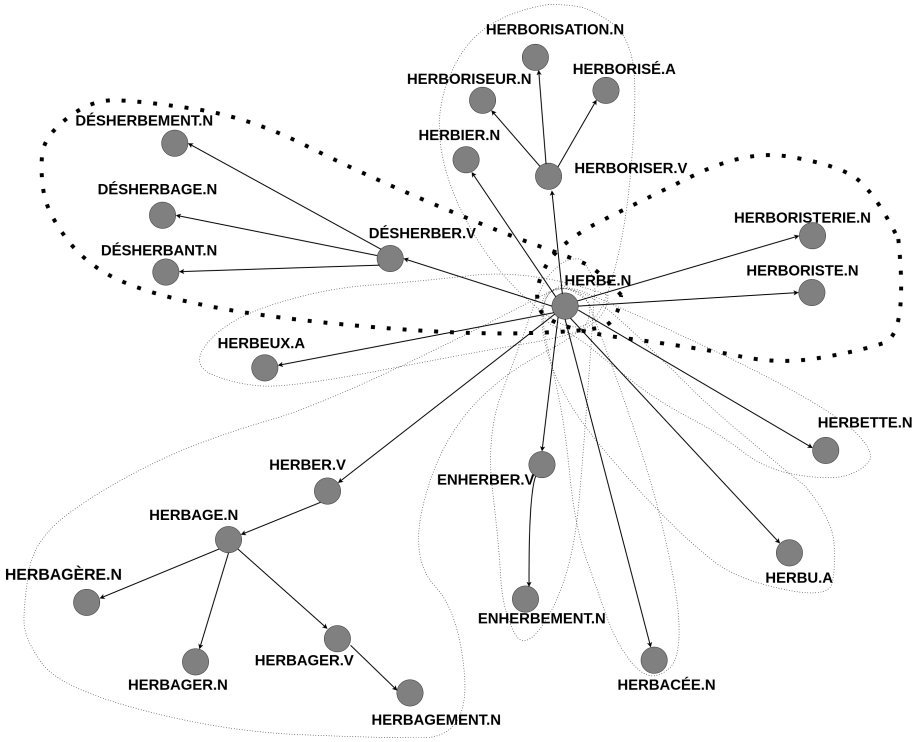

Figure 1: Derivational family of *herbe* 'weed' in French. Edges represent derivational relations. The POS of lexemes are given after the dots. Dotted lines enclose subsets of potentially interpredictable lexemes. We focus on two subsets in particular: the one that is highlighted on the right side contains words related to medicinal herbs selling, while the one on the left side contains words related to weeding.

noun HERBE 'weed' shown in Figure 1; this family is extracted from the French morphological database Démonette (Namer et al. 2023). Semantically, it is hard to argue that all the lexemes in Figure 1 predict one another. For example, the noun DÉSHERBANT 'weedkiller' and the noun HERBORISTERIE 'herbalist's shop' are not interpredictable: the former denotes a chemical used to eliminate invasive weeds, while the latter denotes a store where medicinal herbs are sold. Similarly, the verb ENHERBER 'grass a piece of land' denotes a process by which land is grassed over, while the noun HERBIER 'herbarium' refers to a collection of desiccated plants preserved for scientific purposes. Therefore, the derivational family in Figure 1 can hardly be considered as part of a single derivational paradigm, because its elements are not all semantically interpredictable (Sanacore 2023: 43).

However, we can identify in Figure 1 several subsets containing lexemes that can be considered semantically interpredictable. For example, two distinct subsets are presented in (1). The one in (1a) contains HERBORISTERIE, which is the place where the referent of HERBORISTE works and where medicinal herbs are sold. On the other hand, in (1b), DÉSHERBER, DÉSHERBAGE and DÉSHERBEMENT denote processes of weed removal, and DÉSHERBANT a chemical agent used to carry out this removal. The two subsets in (1) are characterized by a certain "semantic coherence", because they denote concepts that are semantically close and interrelated. Moreover, the subsets in (1) overlap since they both contain HERBE, but they semantically "depend" on two distinct referents of this lexeme: the one in (1a) is related to medicinal herbs, while the one in (1b) is related to invasive wild herbs.

(1) a. HERBE.N, HERBORISTERIE.N ‘herbalist shop’, HERBORISTE.N ‘herbalist’
b. HERBE.N, DÉSHERBER.V ‘to weed’, DÉSHERBAGE.N ‘weed removal’, DÉSHERBANT.N ‘weedkiller’, DÉSHERBEMENT.N ‘weed removal’

On this basis, we consider that derivational paradigms contain partial word families featuring lexemes that are semantically interpredictable. As a consequence, we assume that the subsets in (1a) and (1b) are inscribed in two distinct derivational paradigms. Moreover, we consider that other partial families can be aligned with (1a) and (1b) in their respective derivational paradigms. This hypothesis is similar to what Bonami & Strnadová (2019) proposed in their *paradigmatic systems* and Hathout & Namer (2022) proposed in their paradigmatic model *ParaDis*.

In the last decade, some examples of semantically-based derivational paradigms have been discussed: for example, Roché (2023), Fradin (2020) and Sanacore (2023) discuss the existence of derivational paradigms structured by human manufacturing activities (2a), fruit production (2b) and natural lifecycle of wild animals (2c).

(2) a. POT ‘pot’, POTERIE ‘pottery’, POTIER ‘potter’
BOTTE ‘boot’, BOTTERIE ‘bootmaking’, BOTTIER ‘bootmaker’
b. POMME ‘apple’, POMMIER ‘apple tree’, POMMERAIE ‘apple orchard’
COCO ‘coconut’, COCOTIER ‘coconut tree’, COCOTERAIE ‘coconut plantation’
c. LOUP ‘wolf’, LOUVE ‘female wolf’, LOUVETEAU ‘wolf cub’
LION ‘lion’, LIONNE ‘lioness’, LIONCEAU ‘lion cub’

The objective of the procedure that we propose in this paper is to delimit derivational paradigms on a semantic basis and obtain what Hathout & Namer (2022) call paradigmatic families (i.e. partial word families containing lexemes that belong to the same derivational paradigm) like those presented in (1) and (2)[1]. To do so, we propose to use scenarios that we can obtain starting from short stories. The implementation of this approach is described in Section 4, while Section 3 presents the notions of *scenario* and *frame* more in detail.

3 A scenario-based method

In this paper, we adopt the principle that semantic relations structure derivational paradigms (Štekauer 2014; Bonami & Strnadová 2019; Fernández-Domínguez et al. 2020) and we consider that the delimitation of derivational paradigms should be based on the semantic properties of the lexemes and the relations contained in the derivational lexicon (Sanacore et al. 2021). Furthermore, we believe that the answer to the question of paradigm delimitation depends primarily on the notion of semantic predictability in the constructed lexicon and in derivational families.

On this basis, we propose to define predictability in derivation by means of scenarios. Scenarios are a well-known concept in linguistics (Sanford & Garrod 1998; Erk & Herbelot 2024): they are defined as representations of specific real-world situations (e.g. buying some goods, cooking a meal, teach students). Within the same scenario, participants are closely related on a semantic basis and function as a whole; we consider that this resembles to how paradigms work in morphology[2].

[1] We point out that in this paper we will not deal with the alignment of partial families in the same paradigm.
[2] In a recent work, McNally et al. (2024) also propose to use the notion of scenario to define predictability in derivation.

In our proposal, we also take inspiration from the notion of *frame*, which is close to the notion of scenario (Fillmore 1976; Petruck 1996; Ruppenhofer et al. 2016). A frame is defined as *"a conceptual structure that describes a particular type of situation, object, or event along with its participants and properties"* (Ruppenhofer et al. 2016: p.5). In *FrameNet*[3], a lexical database based on frame semantics (Baker et al. (1998); Ruppenhofer et al. (2016)), frames are presented by short descriptions that introduce the situation that they represent and the prototypical participants involved in that situation. An example of frame description is provided in (3) and concerns a commercial scenario. The participants of the frame in (3) are highlighted in bold : in a prototypical commercial scenario we expect to find a buyer, a seller, some goods that are transferred from the seller to the buyer and some money paid in the exchange. Moreover, the frame description in (3) highlights how the participants are related within the frame: upon agreement, the seller gives to the buyer some goods and receive money in exchange. In addition, in *FrameNet*, the frames descriptions usually feature some definitions that are specific to a given participant and explain how it is related to the other participants of the frame. Two examples of these definitions are provided in (4) for the Buyer and the Money in a commerce scenario.

(3) *COMMERCE_SCENARIO: Commerce is a situation in which a* buyer *and a* seller *have agreed upon an exchange of* money *and* goods *(possibly after a negotiation), and then perform the exchange, optionally carrying it out with various kinds of direct payment or financing or the giving of change. The* seller *indicates their willingness to give the* goods *in their possession to a* buyer *who would give them some amount of* money.

(4) a. buyer: The buyer has the money and wants the goods
b. money: money is given in exchange for goods in a transaction

In the framework of frame semantics, it is generally assumed that frames are culturally-based and are independent of any linguistic realization. However, frames can be instantiated in language use (Petruck 1996; Ruppenhofer et al. 2016). For example, the short sentence in (5) instantiates the frame in (3): the participants involved in the situation described in (5) instantiate the participants of the prototypical commercial scenario described in (3), as it is shown in Table 1: *Abby* is the Buyer, *Robin* is the seller, the *car* plays the role of exchanged goods and *5000 dollars* is the amount of money exchanged.

(5) *Abby bought a car from Robin for 5000 dollars*

Buyer	Seller	Goods	Money
Abby	*Robin*	*car*	*5000 dollars*

Table 1: Correspondence between the participants of the situation described in (5) and the frame prototypical participants presented in (3).

For each frame presented in *FrameNet*, the database presents not only the prototypical participants, but also the words that denote the event that they describe. These words are called *lexical units* in *FrameNet*. For example, in the sentence in (5) the verb *bought* evokes in the mind of the reader a commercial activity and determines the type of event in which the participants intervene. Other lexical units that can be associated with a commercial scenario are *sell*, *purchase* or *commerce*.

Frame semantics and resources like *FrameNet* make the assumption that there is a limited set of conceptual structures (i.e. the frames) that are independent of any linguistic

[3] https://framenet.icsi.berkeley.edu/

structure and that can be instantiated in language use, as we exemplified in (5) and in Table 1. Returning to the initial problem of delimiting derivational paradigms and exploring predictability in derivation, we make a similar assumption and we consider that there is a limited inventory of conceptual structures that pre-exist derivational families and that are relevant for the organization of the derivational lexicon. We believe that the identification of these conceptual structures can thus lead to the identification of semantically-based derivational paradigms. In Section 4 we illustrate how this scenario-based approach can be put into practice using stories and LLMs.

4 Using stories to describe morphosemantic relationships

We propose to characterize the morphosemantic relations contained in derivational families by means of stories that describe such relations in context. Stories enable us to insert morphosemantic relations into situations that involve other participants that are related to the concepts denoted by the lexemes in derivational families. Our proposal is based on the hypothesis that participants that are strongly related from a semantic point of view will frequently co-occur in stories and we will be able to include them in the same prototypical scenario on a distributional basis. For example, if we ask French speakers to tell a short story that contains the words HERBE and HERBORISTERIE, we expect that these stories will regularly involve a shop, some customers and some commercial transactions involving medicinal herbs. On the other hand, if we ask to tell some stories that contain HERBE and DÉSHERBER, we expect that these will regularly involve a weeding activity, some chemical products that are used to realize it and an area that has been invaded by wild herbs. This assumption can also be extended to pairs extracted from other derivational families in French. For example, if we ask for stories built on CLOU 'nail' and CLOUTIER 'nailsmith', we expect that they will describe scenarios where a nailsmith crafts nails in a workshop using some materials in order to sell them to someone else. On the other hand, we expect that stories built on CLOU and CLOUER 'to nail' will describe scenarios where nails are used to repair or build objects.

Concerning the choice of using stories, we consider that producing stories built on pairs of morphologically related words is preferable over using corpus data to estimate semantic predictability because: (a) stories guarantee more control on the context of the morphosemantic relations that we want to describe; (b) stories are text genres where relations between the participants need to be overtly expressed to ensure textual coherence and clarity: this makes it easier to recognize the participants of the situation that they describe and how they are related.

We propose to use short stories (between 75 and 100 words) that are built on a pair of lexemes that belong to the same derivational family [4]. A representative example of the stories that we intend to collect is given in (6). The story is built on HERBORISTE and HERBORISTERIE and features an old local herbalist who sells herbs in her shop to customers needing medical treatment for nausea or similar problems. The main participants are highlighted in bold.

(6) ***La vieille herboriste du quartier*** *tenait une* ***herboristerie*** *depuis cinquante ans. Tous les matins, elle ouvrait sa* ***boutique****, saluant ses* ***clients*** *avec un sourire chaleureux. Un jour, une* ***jeune femme enceinte*** *entra, inquiète pour* ***son futur enfant****.* ***L'herboriste*** *lui prépara un* ***mélange d'herbes*** *pour apaiser ses* ***nausées****. La* ***jeune femme*** *repartit*

[4] We limit the length of stories because the longer a story is, the more participants it will contain who may be distant from the relation that we are trying to describe.

avec reconnaissance, promettant de revenir.
'The **old local herbalist** had been running an **herb shop** for fifty years. Every morning, she opened her **store**, greeting her **customers** with a warm smile. One day, a **young pregnant woman** came in, worried about her **unborn child**. The **herbalist** prepared a **mixture of herbs** to soothe her **nausea**. The **young woman** left gratefully, promising to return.'

The central aspect of stories for predictability estimation is the semantic content that regularly appears in them without being part of the material from which the stories have been created (i.e. the pairs of words). We expect some of the participants that appear in the story in (6) to be rather prototypical in the stories that French native speakers would build on HERBORISTE and HERBORISTERIE: if a herbalist and her shop are involved in the story, there will probably be also some customers, some medicinal herbs that the herbalist sells them and a reason why the customers come to ask the help of the herbalists. On the other hand, the story in (6) does not mention any weed removal processes or any weedkiller, because they would not be "coherent" with the semantic relation linking HERBORISTE and HERBORISTERIE.

In (7) we provide an example of a short story built on HERBE and DÉSHERBER. In this case, the story does not mention any commercial activity, but rather a weeding process involving a gardener who is in charge of fixing an invasive weed problem in a garden. The story also mentions some equipment (presumably tools or products) used by the gardener in the weeding process. We consider that participants like a garden (or any other place where invasive weeds can grow and may need to be removed), a person whose mission is to eliminate invasive weeds and its equipment are prototypical for a scenario involving the words HERBE and DÉSHERBER. On this basis, we expect such participants to frequently occur together in stories built on this pair of words.

(7) *Dans le **petit village de Saint-Pierre**, il y avait un **jardin magnifique** qui ornait la **place du marché**. Mais récemment, une **herbe sauvage invasive** s'était mise à grandir partout, écrasant les **fleurs** et les **arbustes** sous son poids. Les **habitants de Saint-Pierre** étaient dégoûtés par cette situation et demandaient que quelqu'un vienne désherber le **jardin** pour enlever ces **mauvaises herbes**. Le jour vint où un **brave jardinier** arriva avec son **équipement** et commença à désherber le **terrain**. Il travailla durant des heures, élevant soigneusement les **racines de l'herbe envahissante** pour la jeter à l'écart. Enfin, après beaucoup d'efforts, le **jardin** était restauré à sa beauté d'origine.*
In the **small village of Saint-Pierre**, there was a **magnificent garden** that adorned the **market square**. But recently, an **invasive weed** had started to grow everywhere, crushing **flowers** and **shrubs** under its weight. The **people of Saint-Pierre** were disgusted by this situation and demanded that someone come and weed the **garden** to remove the **weeds**. The day came when a **brave gardener** arrived with his **equipment** and began to weed the **grounds**. He worked for hours, carefully pulling up the **roots** of the **overgrown weed** and tossing it aside. Finally, after much effort, the **garden** was restored to its original beauty.

The procedure that we propose to obtain derivational paradigms using stories is schematized in the flowchart in Figure 2. In the remainder of this Section, we describe each of the four steps it involves. In Subsection 4.1, we describe how we generate stories from pairs of words extracted from a derivational family using LLMs (we take the family of HERBE presented in Section 2 as representative example). In Subsection 4.2 we describe the manual annotation of stories, in which we assess the instantiation of prototypical

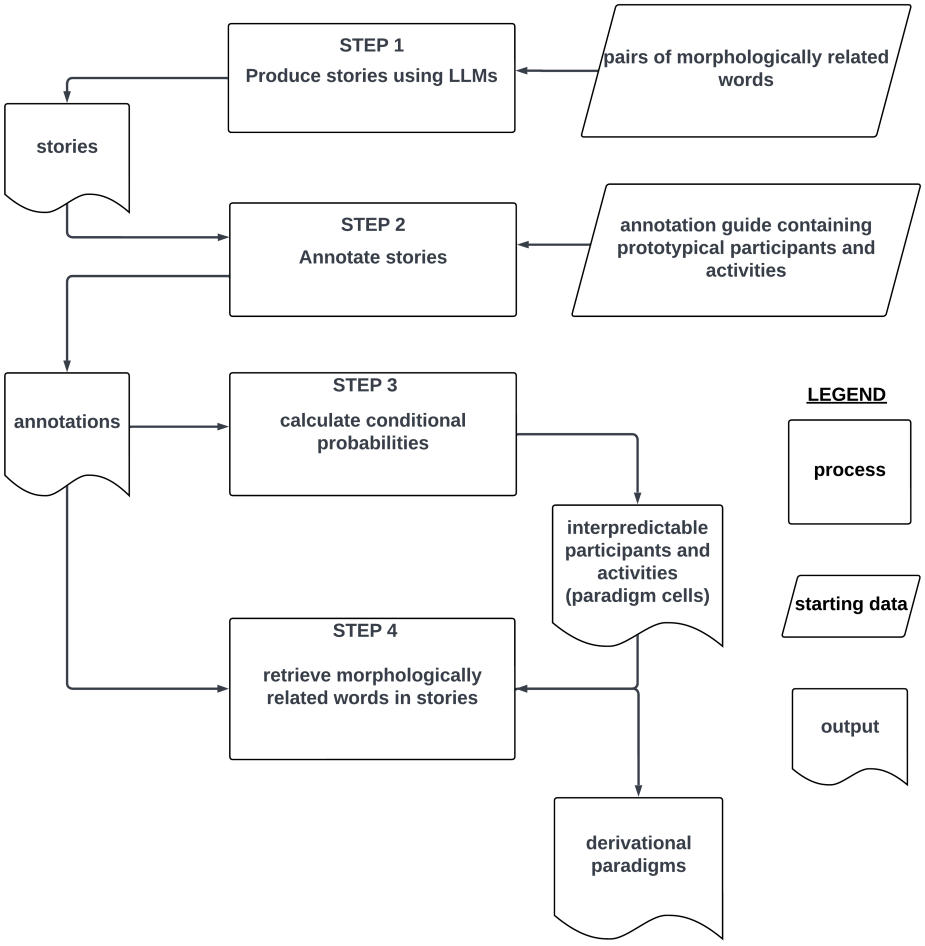

Figure 2: Diagram representation of the procedure for the identification of derivational paradigms using stories. Each one of the four steps is detailed in the remainder of this section.

participants and activities in stories. In Subsection 4.3, we describe how we calculate conditional probabilities given the annotation that have been realized in the second step and we identify the participants and activities that are interpredictable in stories. Finally, in Subsection 4.4, we describe how the interpredictable cells identified through conditional probability can be filled to reconstruct the derivational paradigms.

4.1 Generating stories with LLMs

In order to collect a large quantity of stories to identify predictable semantic relations, we propose to generate them using Large Language Models (LLMs). We consider that LLMs are relevant for this task since they are capable of producing large quantity of semantically coherent texts in a relatively short time and they are capable of producing stories that reflect the real world knowledge that a human speaker could have. For example, the two stories presented in the previous section in (6) and (7) have been automatically produced by Llama3 [5], a language model trained using data in several languages, including French.

[5] https://ai.meta.com/blog/meta-llama-3/

To test the relevance of LLMs for semantic predictability estimation and the validity of our scenario-based hypothesis, we realized a preliminary story production task using *Ollama* [6], a free and open source tool that enables users to use LLMs locally on a personal computer without relying on web-based interfaces like *ChatGPT* or *Gemini*. Using LLMs locally on our machine enables us to: (a) avoid limitations in terms of runtime or output size; (b) automatically save the output in a locally stored file and fully automate the story production step. On the other hand, using LLMs locally on a personal machine with *Ollama* means relying on limited hardware and have access to less powerful models compared to the ones that can be accessed via the *ChatGPT* and *Gemini* web-based interfaces. For this story production task, we used the 8B parameters version of Llama3, one of the mostly capable freely available LLMs, which proved to be sufficiently powerful to produce realistic and semantically coherent stories in French.

We realized our story production task taking the family of HERBE presented in Section 2 as example to test our approach. Using a Python notebook, we asked Llama3 to produce 400 short stories starting from 8 pairs extracted from this derivational family. The pairs are given in (8) [7]. Among them (8), we can distinguish: three pairs of lexemes linked by a morphosemantic relation that we assume to be inscribed in a weeding scenario (8a), three pairs of lexemes linked by a relation that we assume to be inscribed in medicinal herbs commerce scenario (8b) and two pairs of lexemes whose semantic relation is not straightforward (8c). We consider that the pairs in (8c) do not constrain the content of the story and we expect that the stories produced by the LLMs from these pairs will involve medicinal herbs commerce, weeding, or both.

(8) a. HERBE-DÉSHERBER; DÉSHERBER-DÉSHERBANT; HERBE-DÉSHERBANT
b. HERBE-HERBORISTE; HERBORISTE-HERBORISTERIE; HERBE-HERBORISTERIE
c. DÉSHERBER-HERBORISTE; DÉSHERBANT-HERBORISTERIE

A central aspect of any interaction with generative tools such as LLMs is the prompt (i.e. what we actually ask the language model to do). The prompt used to make the model produce the stories is given in (9). A first aspect to point out is that we formulated our prompt in English even though we were producing French data: this is because Llama3 and practically all the other available large language models are mostly trained on English, which therefore is the language in which they are more proficient. For this reason, formulating our prompt in English granted the highest chance of avoiding any misunderstanding with the language model and have it doing precisely what we asked. Moreover, we asked the model to produce short realistic stories that reflect the real world knowledge of speakers, in order to limit the tendency of LLMs to produce fable-like stories involving humanized animal participants or any other non-realistic fact. Lastly, we asked the model not to produce stories in first person in order to make it "declare" more overtly the participants involved in the story and make them more recognizable.

(9) *" You are a French native speaker. You will always answer in French, without using English. You will never provide an English translation of your answer. Your answers will be stories containing no less than 75 words and no more than 100 words. The stories that you will produce will be realistic, not fables. If the story involves animals, it*

[6] https://ollama.com/

[7] The code, the stories produced by Llama3 and their annotation can be found at the following URL: https://gitlab.com/llcd-sanacore-hathout-namer/llcd-histoires-sanacore/-/tree/main/Histoire-herbe-septembre24?ref_type=heads

will be realistic with respect to what animals do in the real world. You will never make stories in first person."

Using the instruction in (9) as system prompt, for each one of the pairs given in (8), we asked the model to produce a story using the request template in (10). The request template in (10) and the system prompt in (9) can be used to produce stories starting from any pair of morphologically related words.

Together with the pairs of morphologically related words in (8), we gave the model their part of speech and a short gloss expressing the semantic relation between the two words (when the semantic relation was straightforward). The LLM thus knows the relation between *word1* and *word2* before producing the stories, in the same way as a native speaker would know the relation between the two words before starting telling a story. Moreover, using glosses to describe semantic relations helps to avoid any misunderstanding with the model on the semantic properties of the lexemes that we are using. This can be particularly crucial when dealing with relations that involve polysemous or rarely attested words. For the two pairs containing words that are not semantically related (8c), we simply asked the model to produce stories that contained the two concerned words. We highlight that the objective of the story production task is not to discover the relation between *word1* and *word2*, but rather to see which relations regularly appear in the same story as the relation between these two words.

The information given as input to the language model is provided in Table 2 and the instruction presented in (10) has been reiterated 50 times for each pair of words, producing 400 stories in total. Moreover, the language model has been reinitialized at each iteration of the algorithm in order to decrease the probability of repetitions in stories.

(10) *"Tell me a story that contains the words* WORD1 *(which is a* CAT1*) and* WORD2 *(which is a* CAT2*). The story should reflect the following relationship between these words:* RELATION.[8]*"*

4.2 Semantic role annotation in stories

As mentioned at the beginning of Section 3, our approach is based on the assumption that there is a limited set of scenarios that are independent of language use (similarly to frames) and that are potentially relevant for the structure of the derivational lexicon. Moreover, we consider that for each scenario, there is a limited number of prototypical participants (e.g. the buyer, the instrument used to eliminate an harmful entity, the cultivated field, etc.) and activities (buying, farming plants, etc.).

After that the stories have been automatically produced, we propose to realize an annotation task in order to assess the instantiation of a set of prototypical activities and participants in stories, in order to empirically validate them and assess their indipendence from each other. These prototypical activities and scenarios are thus part of the starting data and are used for the annotation: the objective of stories is to empirically evaluate their instantiation in stories and retrieve the participants and activities that are interpredictable. For example, if a buyer, a seller and a commercial activity are systematically instantiated in the same story, but they are almost never instantiated together with an elimination activity, this means that the elimination activity belongs to a different "package" of interpredictable relations and, consequently, to a different scenario. On the other hand, if stories systematically contain the instantiation of a commercial activity, a

[8] When the description of the relation was not available, we omitted the last sentence of the instruction.

word1	word2	cat1	cat2	relation
herbe	désherber	noun	verb	*désherber signifie débarasser un lieu de mauvaises herbes*
désherber	désherbant	verb	noun	*le désherbant est un produit utilisé pour désherber un terrain ou un jardin*
désherbant	herbe	noun	noun	*le désherbant est un produit utilisé pour éliminer les mauvaises herbes*
herbe	herboriste	noun	noun	*un herboriste est une personne qui vend des herbes et des graines médicinales*
herboriste	herboristerie	noun	noun	*une herboristerie est une boutique dans laquelle travaille un herboriste*
herboristerie	herbe	noun	noun	*une herboristerie est une boutique dans laquelle on vend des herbes médicinales*
désherber	herboriste	verb	noun	-
désherbant	herboristerie	noun	noun	-

Table 2: Input information given to the model to produce the stories. For each pair, the model produced 50 stories.

buyer and an elimination activity together, this means that the commercial activity and the elimination activity could be part of one same prototypical scenario.

Contrarily to the story production step, which has been realized automatically, this step has been realized manually [9]. The participants and activities that we used for the annotation are provided in Table 3, but we plan to extend this inventory in future studies. These labels are inscribed in four hypothetical scenarios that we want to assess: (a) commercial scenario; (b) entity removal; (c) animal lifecycle; (d) plant and fruit cultivation. We selected these four scenarios since they have all been shown as relevant for the paradigmatic organization of the lexicon: the first two are those that we assume to be instantiated in the family of HERBE in our starting hypothesis in Section 2, while the third and the fourth have been discussed relatively to the derivational paradigms cited in (2) at the end of Section 2. The hypothetical scenarios that we chose are sufficiently general in order to be instantiated by a set of distinct stories produced from distinct morphologically related pairs but, at the same time, they enable to distinct real-world situations that are different in nature (e.g. commerce *vs* entity elimination).

For a commercial activity, we considered that there typically is a seller, some goods that are sold, a buyer who buys them and a shop where the commercial activity takes place. For an elimination activity, we considered that there usually is a human eliminator who eliminates an commercial eliminated entity from a liberated place and that the eliminator may use either an instrument or some means to realize this activity. The distinction between instrument and means is based on the reusability constraint proposed by Fradin & Winterstein (2012). For the animal lifecycle, we considered that the prototypical participants are male adult a female adult and an offspring of a wild animal species. Lastly, we considered that in a plant cultivation activity the prototypical participants are: a farmer

[9] In this work, the annotation has been realized by the first author of this paper. Its objective is to contribute to the fine-tuning of the methodology that we propose, while it is not aimed at creating a reference resource.

that cultivates the plant, the cultivated plant, a cultivation site where the plant grows and some cultivated fruits that grow on this plant.

Since the purpose of the story production task is to empirically test the four aforementioned hypothetical scenarios by identifying the groups of activities and participants that are highly interpredictable, the labels selected for the annotation task are presented all together in the annotation guide, rather than sorted according to the scenarios into which we hypothesize they can be grouped. In the annotation guide, for each participant and activity, we provided a definition that helps the annotator identifying it in stories. For space reasons, the labels and definitions are directly presented in English.

Participants and activities	**Definition**
seller	person selling a property to a buyer
shop	place where a commercial activity is realized
merchandise	item sold from buyer to seller
customer	person who buys a good
commercial activity	activity in which one person sells a good to another
eliminator	person who removes an entity from an infested area
eliminated entity	entity removed from an infested area
eliminating means	product used to remove an entity from a given area (it is modified during the process)
eliminating instrument	artifact used to remove an entity from a given area (it is not modified during the process)
liberated place	natural or artificial zone freed of a given harmful entity
elimination activity	activity in which one person removes an entity from a given area
wild animal-male	male specimen of a wild animal species
wild animal-female	female specimen of a wild animal species
wild animal-offspring	offspring of a wild animal species
farmer	person who grows a plant
cultivation site	place where plants are cultivated
cultivated plant	plant cultivated by one or more people
cultivated fruit	fruit produced by a cultivated plant
farming activity	activity in which plants are cultivated

Table 3: Label set used for the annotation. The prototypical participants and activities presented in this table can be used for the annotation of stories produced starting from any derivational family in French.

To provide a story annotation example using the participants and activities in Table 3, in the story in (11), the annotator is supposed to mark the presence of two activities: *Monsieur Leblanc* both cultivates plants and vegetables in his garden and removes the wild weeds that invaded it. This example shows that a story element can instantiate two prototypical participants: *Monsieur Leblanc* is both the farmer with respect of his flowers and vegetables and the eliminator of the invasive weeds. Moreover, the garden is both the cultivation site where plants grow and the liberated place of the invasive weeds. The full annotation for the story in (11) is shown in Table 4. When a prototypical participant of the annotation guide is not instantiated in a story, the annotator marks '-' in the correspondent cell.

Protot. participants and activities	Story element
eliminator	*Monsieur Leblanc*
eliminated entity	*mauvaises herbes*
eliminating means	*désherbant*
liberated place	*jardin*
elimination activity	*mettre fin à l'invasion végétale*
farmer	*Monsieur Leblanc*
cultivated plant	*fleurs, légumes, herbes*
cultivated fruit	*tomates*
cultivation site	*jardin*
farming activity	*cultiver*
...	-

Table 4: Annotation of the story in (11). The last line in the table represents all the prototypical participants and activities of the annnotation set that are not instantiated in this story.

(11) *La **propriété** de **Monsieur Leblanc** était entourée d'un **jardin verdoyant**, où il aimait passer des heures à cueillir les **légumes et les fleurs qu'il cultivait lui-même**. Mais voilà que les **mauvaises herbes** s'étaient mises à prendre le dessus, envahissant les rangs de **tomates** et d'**herbes fraîches**. **Monsieur Leblanc** décida donc d'utiliser un **désherbant** pour mettre fin à cette **invasion végétale** [...].*
'**Monsieur Leblanc**'s **property** was surrounded by a **verdant garden**, where he enjoyed spending hours picking the **vegetables and flowers** he grew himself. But **weeds** had taken over, invading the rows of **tomatoes** and **fresh herbs**. **Mr. Leblanc** decided to use a **weedkiller** to put an end to this **plant invasion** [...].'

The result of the annotation of the 400 stories is a table where, for each story, the annotator marked the prototypical participants and activities that are instantiated and the syntactic element of the story that instantiates it.

4.3 Calculating conditional probability in stories

Once all the stories have been annotated, we computed conditional probability in order to assess the interpredictability of the prototypical participants that we have used for the annotation. The conditional probability of an event B is the probability that the event will occur given the knowledge that an event A has already occurred. The formula is given in (12).

(12)
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{with} \quad P(B) > 0$$

We are interested in two types of conditional probability. Firstly, we computed the conditional probability of a participant B (e.g. buyer) to be realized in a story created starting from a pair of derivationnally related lexemes A (e.g. HERBE-DÉSHERBER). The results are shown in the heatmap in Figure 3. The stories produced from the first three pairs systematically describe an elimination of an invasive entity from a place (mostly invasive weeds). Within these stories, the instrument and the means used for the elimination are in a complementary distribution: stories built on a pair including DÉSHERBANT

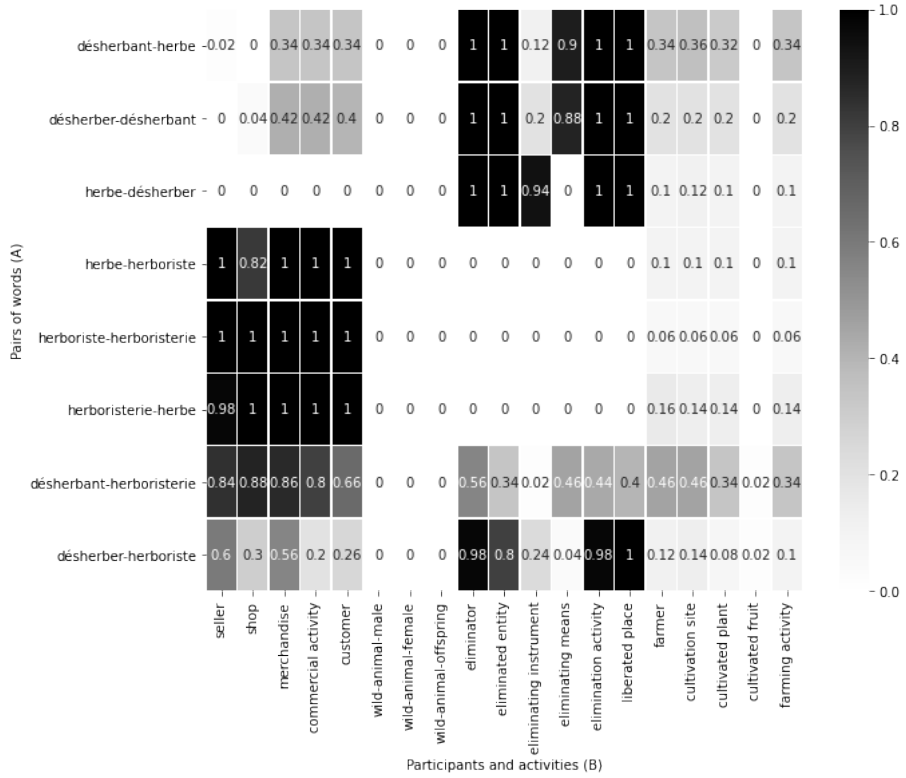

Figure 3: Conditional probabilities of a prototypical participant B (e.g. customer) to be instantiated in a story produced starting from a word pair A (e.g. DÉSHERBER-DÉSHERBANT)

mostly involve means (i.e. weedkillers), while stories built on the HERBE-DÉSHERBER pair mostly involve instruments like rakes or electric lawnmowers and rarely mention chemical products used for weeding. Moreover, stories built on the first three pairs in Figure 3 sometimes mention the purchase of the weed killer by a gardener or a farmer (but almost never mention the seller and the shop) and sometimes that the liberated place is a cultivated field. On the other hand, stories built on the pairs from the fourth to the sixth row systematically describe commercial activities and never mention the elimination of an entity from a given area. In some cases, in these stories, the products of the herbalist shop are cultivated by the herbalist in a garden or a field.

As we expected, the stories built on the seventh and eight pair in Figure 3 are "mixed" in terms of the scenario that they describe: stories built on the pair DÉSHERBANT-HERBORISTERIE usually mention some herbalist shops that also sell weedkillers to be used in cultivated fields, while stories built on DÉSHERBER-HERBORISTE usually tell about herbalists using weedkillers in their own cultivated field. Lastly, none of the 400 stories produced mentioned wild animal specimens or any relations between them.

However, the heatmap shown in Figure 3 does not show us which cells are predictable one from the other, but rather which cells are predictable given the pair of words from which the stories have been created. Moreover, it can be misleading on some probabilities: for example, the fact that the stories produced from the pair DÉSHERBER-HERBORISTE often contain elimination activities does not mean that elimination activities and herbalists are highly interpredictable, but rather that the model produced stories that fit the constraint of contaning both the words HERBORISTE and DÉSHERBER and that tell

about herbalists that also do some weeding or some herbalists that sell weedkillers. The same problem applies to DÉSHERBANT-HERBORISTE.

For this reason, in order to have a better understanding of predictability relations between participants in our stories, we computed the conditional probability of a participant or activity C to be realized in a story given the presence of another participant or activity B in the same story. The resulting heatmap is provided in Figure 3. Three high interpredictability areas (i.e. where probabilities are more than 0.75 in both directions) can be identified: in the top left corner, we can see that when a commercial activity occurs in a story, there will probably be a seller, a customer, a merchandise and a shop, but the probability to have a cultivation activity or some eliminated entities is quite low. Conversely, given the presence of an elimination activity in a story, we systematically find in the same story an eliminator and a liberated place. The instrument and the means are in complementary distribution and their conditional probabilities given the presence of an elimination activity are lower, but the probability to have either one or the other in a story with an elimination activity or an eliminator is higher than 0.80. Finally, a third high interpredictability area can be identified in the bottom right corner and corresponds to the participants and activities that feature in a cultivation scenario: a farming activity, a farmer, a cultivation site and a cultivated plant[10]. We point out that the results also show that the network built around the elimination activity and the one built around a farming activity cannot be considered as interpredictable: for example, given the presence of a farming activity in a story, the probability of having an elimination activity is close to 0.75, but the probability of having a farming activity given the presence of elimination activity is much lower.

To sum up, the heatmap in Figure 3 enables us to identify three distinct semantic networks that fit in three distinct scenarios. It also shows that these networks are not predictable one from another and confirms our hypothesis that the derivational family of HERBE, from which we started, is structured by distinct semantic networks that may overlap. This means that we potentially have (at least) three distinct derivational paradigms structured by three distinct scenarios.

4.4 Filling the derivational paradigms

The aim of the last step of the procedure that we propose is to see which cells of the three networks identified in Figure 4 are realized in stories by morphologically related words and identify the derivational paradigms that we are looking for. The manner in which this step will be carried out is yet to be fine-tuned, but the objective is to realize it automatically using the story annotations produced in the second step and the interpredictability areas identified in the third phase, which enabled us to identify the cells to be filled.

For example, in the story annotations, we can see that HERBORISTE often instantiates the seller in stories, HERBORISTERIE the shop and HERBE the merchandise. For what concerns the elimination scenario, DESHERBER and DÉSHERBAGE instantiate the elimination activity, HERBE the eliminated entity, DÉSHERBANT the eliminating means and DÉSHERBEUR[11] the eliminating instrument. Moreover, in some stories, the model *LLama3* used DÉSHERBEUR to refer to the eliminator .

[10] The cultivated fruit has been omitted from this second heatmap because it is only instantiated in two stories in the whole dataset.

[11] DÉSHERBEUR is not attested in the derivational database *Démonette*, but can be easily encountered in many web pages that deal with gardening.

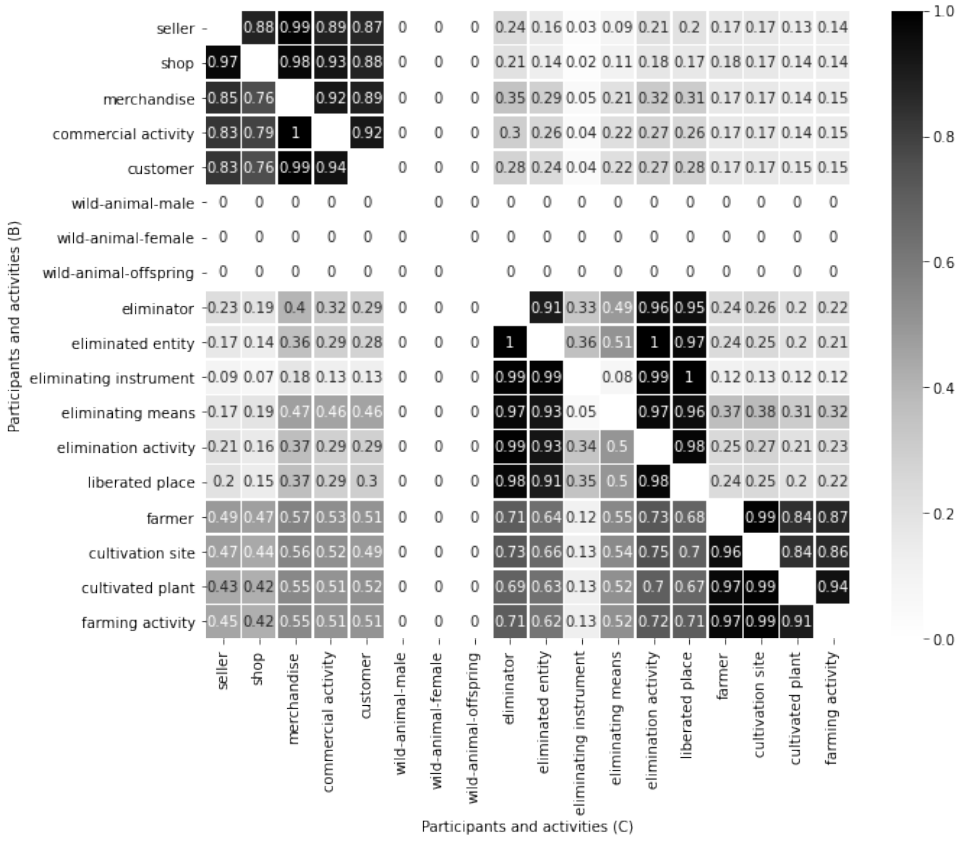

Figure 4: Conditional probabilities of a participant or activity C (e.g. SELLER) to be instantiated in the same story of a participant or activity B (commercial activity).

This last step thus enables us to project the semantic networks identified through stories onto morphology and see which participants are realized by morphologically related lexemes. Concerning the family of HERBE, from which we started, the two distinct paradigmatic families that we can extract using story annotations are given in (13): they confirm what we had hypothesized at the beginning of the paper in (1) in Section 2. For what concerns the third network structuring a farming activity that we identified in Figure 4, a third paradigm that can be extracted from this family is suggested in (13c). In this third paradigm, HERBE instantiates the cultivated plant cell. The paradigm also features ENHERBER as grass farming activity and ENHERBEUR, which denotes a tractor used for seeding grass. The presence of ENHERBEUR suggests that the possibility to add a new prototypical participant that we could call farming instrument should be investigated using stories.

(13) a. HERBE; DÉSHERBER; DÉSHERBAGE; DÉSHERBEUR; DÉSHERBANT
b. HERBE; HERBORISTERIE; HERBORISTE
c. HERBE; ENHERBER; ENHERBEUR

5 Conclusions and future perspectives

In this study, we proposed a scenario-based methodology in order to explore the notion of predictability in the constructed lexicon. More precisely, we proposed to produce a large quantity of stories starting from pairs extracted from the same derivational family

in order to identify the concepts that are semantically predictable given the relation between *word1* and *word2* and assess interpredictability in derivational families.

We realized a first empirical task using LLMs to produce stories starting from the family of HERBE in French. However, the procedure that we propose can be reproduced for any derivational family. Once the stories have been automatically produced, we realized a manual annotation following an annotation guide where we organized prototypical activities and participants of four scenarios that we wanted to empirically validate using stories. The annotation objective is to assess the instantiation of the prototypical participants and activities in the stories produced by the model, in order to see which participants and activities are regularly instantiated together.

Once the annotation task has been realized, we calculated two conditional probabilities: (a) the conditional probability of a participant or activity B to be instantiated in stories produced from a word pair A (b) the conditional probability of a participant or activity C to be instantiated in a story, given the presence of a participant B. The results of the calculation of the conditional probability between participants show that there are three distinct semantic networks that are independent and not interpredictable: the first structures a commercial activity, the second structures an activity where one eliminates an invasive entity and the third structures a plant farming activity. This validates the scenarios that we had hypothesized and reveals that there are (at least) three distinct derivational paradigms that can be identified in the family of HERBE. The last stage of the procedure consists in finding the participants of the three networks that are realized by morphologically related lexemes and enables to identify the paradigmatic families in (13).

In future work, we intend to apply the procedure that we proposed on a large quantity of French derivational families using *Démonette* (Namer et al. 2023). Using stories to describe relations inscribed in a large quantity of derivational families, we intend to explore not only the possibility to "slice" derivational families into paradigmatic families, but also the possibility to take paradigmatic families extracted from different derivational families and align them in the same derivational paradigm.

We also intend to do further work on the procedure that we have proposed in this paper. Firstly, we would like to extend the range of prototypical participants and activities to use in the annotation guide. Drawing from the frame dataset in *FrameNet* could be helpful on this point. Secondly, we intend to explore and evaluate new ways of prompting LLMs, in order to see which prompts are more efficient for story generation tasks. Lastly, we intend to test LLMs capacity to realize annotations like the one that we propose and to automatically label participants and activities in stories. Training LLMs to efficiently realize the participant labeling task could enable us to fully automate our procedure.

References

Baker, Collin F & Fillmore, Charles J & Lowe, John B. 1998. The berkeley framenet project. In *Coling 1998 volume 1: The 17th international conference on computational linguistics*.

Bauer, Laurie. 2019. Notions of paradigm and their value in word-formation. *Word Structure* 12(2). 153–175.

Beniamine, Sacha. 2018. *Classifications flexionnelles. étude quantitative des structures de paradigmes*: Université Paris Cité Thèse de doctorat.

Bonami, Olivier. 2014. *La structure fine des paradigmes de flexion*: Université Paris 7 HDR.

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

Boyé, Gilles & Schalchli, Gauvain. 2016. The status of paradigms. In *The cambridge handbook of morphology*, 206–234. Cambridge University Press.

Dokulil, Miloš. 1982. Kotázce slovnědruhovỳch převodů a přechodů, zvl. transpozice. *Slovo a slovesnost* 43(4). 257–271.

Erk, Katrin & Herbelot, Aurélie. 2024. How to marry a star: Probabilistic constraints for meaning in context. *Journal of Semantics* .

Fernández-Domínguez, Jesús & Bagasheva, Alexandra & Clares, Cristina Lara. 2020. *Paradigmatic relations in word formation*. Brill.

Fillmore, Charles J. 1976. Frame semantics and the nature of language. In *Annals of the new york academy of sciences: Conference on the origin and development of language and speech*. 20–32. New York.

Fradin, Bernard. 2020. Characterizing derivational paradigms. In *Paradigmatic relations in word formation*, 49–84. Brill.

Fradin, BERNARD & Winterstein, Gregoire. 2012. Tuning agentivity and instrumentality: deverbal nouns in-oir revisited. *Paper delivered at Décembrettes* 8. 6–7.

Gaeta, Livio. 2022. Dangerous liaisons. An introduction to derivational paradigms. In *Paradigms in word formation*, 3–18. John Benjamins.

Hathout, Nabil & Namer, Fiammetta. 2019. Paradigms in word formation: what are we up to? *Morphology* 29(2). 153–165.

Hathout, Nabil & Namer, Fiammetta. 2022. ParaDis: a family and paradigm model. *Morphology* 1–43.

McNally, Louise & Bonami, Olivier & Paperno, Denis. 2024. It's time for a complete theory of partial predictability in language. *Theoretical Linguistics* .

Namer, Fiammetta & Hathout, Nabil & Amiot, Dany & Barque, Lucie & Bonami, Olivier & Boyé, Gilles & Calderone, Basilio & Cattini, Julie & Dal, Georgette & Delaporte, Alexander & Duboisdindien, Guillaume & Falaise, Achille & Grabar, Natalia & Haas, Pauline & Henry, Frédérique & Huguin, Mathilde & Juniarta, Nyoman & Liégeois, Loïc & Lignon, Stéphanie & Macchi, Lucie & Manucharian, Grigoriy & Masson, Caroline & Montermini, Fabio & Okinina, Nadejda & Sajous, Franck & Sanacore, Daniele & Tran, Mai Thi & Thuilier, Juliette & Toussaint, Yannick & Tribout, Delphine. 2023. Démonette-2, a derivational database for French with broad lexical coverage and fine-grained morphological descriptions. *Lexique* 33. 6–40.

Petruck, Miriam. 1996. Frame semantics. In Verschueren, J. & Östman, J.-O. & Blommaert, J. & Bulcaen, C. (eds.), *Handbook of pragmatics*, 1–13. John Benjamins.

Pounder, Amanda. 2000. *Process and paradigms in word-formation morphology*. Mouton de Gruyter.

Roché, Michel. 2023. Les familles dérivationnelles: comment ça marche? *Lexique. Revue en Sciences du Langage* 33.

Ruppenhofer, Josef & Ellsworth, Michael & Schwarzer-Petruck, Myriam & Johnson, Christopher R & Scheffczyk, Jan. 2016. Framenet II: Extended theory and practice. Tech. rep. The Berkeley FrameNet Project.

Sanacore, Daniele. 2023. *Une histoire de famille: description morphosémantique des lexèmes construits et des relations dérivationnelles*: Université Toulouse le Mirail-Toulouse II dissertation.

Sanacore, Daniele & Hathout, Nabil & Namer, Fiammetta. 2021. Frame-like structure for morphosemantic description. *Verbum (Presses Universitaires de Nancy)* 43(1). 179–194.

Sanford, Anthony J & Garrod, Simon C. 1998. The role of scenario mapping in text comprehension. *Discourse Processes* 26(2-3). 159–190.

Štekauer, Pavol. 2014. Derivational paradigms. In *The oxford handbook of derivational morphology*, 354–369. Oxford University Press.

Stump, Gregory T. 1991. A paradigm-based theory of morphosemantic mismatches. *Language* 675–725.