Annotating the French *Wiktionary* with supersenses for large scale lexical analysis: a use case to assess form-meaning relationships within the nominal lexicon

Nicolas Angleraud
LLF
Université Paris Cité, CNRS
nicolas.angleraud@gmail.com

Lucie Barque
USPN, Villetaneuse, France
LLF, CNRS, Paris, France
lucie.barque@univ-paris13.fr

Marie Candito
LLF
Université Paris Cité, CNRS
marie.candito@u-paris.fr

Abstract

Many languages lack broad-coverage, semantically annotated lexical resources, which limits empirical research on lexical semantics for these languages. In this paper, we report on how we automatically enriched the French Wiktionary with general semantic classes, known as supersenses, using a limited amount of manually annotated data. We trained a classifier combining sense definition classification and sense exemplars classification. The resulting resource, with an evaluated supersense accuracy of nearly 85% (92% for hypersenses), is used in a case study illustrating how such an semantically enriched resource can be leveraged to empirically test linguistic hypotheses about the lexicon, on a large scale[1].

1 Introduction

Conducting large-scale empirical studies in lexical semantics remains an elusive goal for many languages that lack comprehensive semantic resources. Thanks to its hierarchical organization facilitating semantic generalizations, the Princeton WordNet (Miller et al., 1990), a widely used lexical resource for English, has enabled quantitative studies on regular polysemy (Buitelaar, 1998; Bol) or on morpho-semantic relations (Mititelu et al., 2021), among others. Such studies are difficult to conduct for languages without comparable resources, limiting typological perspectives on key issues in lexical semantics.

In this study, we adopt the intermediate scenario of a language without a sufficiently comprehensive WordNet, but with a comprehensive lexicon and a few thousand senses from this lexicon, annotated into coarse semantic classes, allowing a classifier to be trained to automatically annotate the entire lexicon. We apply this approach to French nouns, producing a semantically enriched version of the French Wiktionary. The method can be extended to other languages provided an electronic lexicon is available (e.g. Wiktionaries[2]) and manually annotated senses. We qualify this scenario as intermediate because the annotation effort required (for training the classifier) is far less compared to the effort needed to obtain a comprehensive WordNet. Our study further demonstrates how such an enriched lexicon can be leveraged to conduct quantitative studies, by empirically testing Croft's hypothesis on the general organization of word and semantic classes within language lexicons (Croft, 1991).

The paper is organized as follows. Section 2 briefly reviews previous work on the development and main uses of broad coverage, coarse-grained semantic resources in NLP and linguistic research. Section 3 describes and evaluates the supervised method for annotating noun senses with supersenses in the French *Wiktionary*. Section 4 provides key statistics of the resulting resource, we dub the "SuperWikt-fr". Section 5 presents a case study to evaluate the extent to which Croft's hypotheses on the distribution of semantic classes across the simplex and complex lexicon are supported by French data. Finally, Section 6 offers suggestions for further exploration and exploitation of semantically enriched lexical resources in linguistic research.

2 Related work

Supersenses are coarse-grained semantic classes originally proposed to help downstream tasks requiring semantic information, for English (Ciaramita and Johnson, 2003). More precisely the original supersenses were the 26 nominal "lexicographer classes" from WordNet. Ciaramita and Altun (2006) then proposed the supersense tagging

[1] The SuperWikt-fr is available at https://osf.io/7gjem/?view_only=42190678aba442b39664ad05a54bf843, and code to reproduce the annotation process on new dumps is available at https://github.com/NicolasAngleraud/SuperWikt-fr.

[2] According to https://en.wikipedia.org/wiki/Wikipedia:List_of_Wiktionaries, 45 languages have a Wiktionary with more than 100 thousand entries.

Proceedings of the 31st International Conference on Computational Linguistics, pages 5321–5332
January 19–24, 2025. ©2025 Association for Computational Linguistics

task (SST) as a trade-off between NER and WSD: it provides partial semantic disambiguation among senses with different supersenses, while being more easily achievable than full WSD. The methods for SST followed the technical evolution of NLP (see e.g. the DIMSUM shared task (Schneider et al., 2016)).

This led to the development of supersense-annotated corpora for several languages, with strategies depending on the availability of semantic resources: directly deriving supersenses from annotated WordNet synsets (for English (Ciaramita and Altun, 2006)), specifying pre-existing annotations (for Arabic (Schneider et al., 2012), Italian (Dei Rossi et al., 2011)), annotating both word senses and their supersense (for Danish (Pedersen et al., 2016)), or annotating supersenses only: e.g. for French, the existing WordNet-like resources (Vossen, 1998; Fiser and Sagot, 2015) did not meet the full requirements of availability, manual quality, and coverage, which led to the manual supersense annotation of the FrSemCor corpus, from scratch (Barque et al., 2020). The semantically-enriched lexicon described in this article aims to provide a counterpart lexical resource for French.

Coarse-grained semantic classes also played a role in the development of distributional semantic representations, first for static embeddings (e.g. Flekova and Gurevych (2016) integrate vectors of words, supersenses, and word-supersense pairs within the same vector space, for English) and then for contextualized representations. The SenseBERT bidirectional model provides contextual representations of both tokens and supersenses (Levine et al., 2020). It is pre-trained on the tasks of predicting a word and a supersense in context. Interestingly, SenseBERT does not use a supersense-annotated corpus, but only a lexicon providing the possible supersenses of a word. The pre-training loss favors the set of possible supersenses of the current word. The authors show that SenseBERT improves on some of the GLUE benchmark tasks.

The advent of large language models seriously calls into question the usefulness of supersense tagging for NLP. The situation is quite different for quantitative lexical studies. While the existence of WordNet has enabled numerous studies for English (such as studies on regular polysemy (Buitelaar, 1998; Peters and Peters, 2000, a.o)), for most languages, questions of lexical semantics cannot today be investigated for lack of a lexicon having both large coverage and semantic classification. This is the case for French, for which empirical studies on the semantic properties of the lexicon are necessarily limited in scope. For example, Tribout et al. (2014) conducted a morpho-semantic analysis of a manually selected set of 3,500 French nouns. The goal was to evaluate some of the linguistic generalizations proposed by Croft (1991) regarding the prototypical correlations between morpho-syntactic categories, semantic classes, and pragmatic functions. According to Croft, a noun (respectively, a verb or adjective) typically denotes an object (respectively, an action or property), with its pragmatic function being reference (respectively, predication or modification). Croft further hypothesizes that a noun with prototypical semantic class (object) would prototypically be simplex, i.e., not morphologically constructed (e.g., *vehicle*), whereas action or property nouns would typically be suffixed (e.g., *destruction* or *whiteness*). Tribout et al. (2014) tested Croft's hypothesis for French by manually assessing the proportion of counterexamples among simplex nouns, namely those denoting an action (e.g., *crime* 'crime') or a property (e.g., *courage* 'courage'). Their findings support Croft's hypothesis that simplex nouns predominantly denote objects but offer a more nuanced view of the semantics of this morphological word class, by notably highlighting non-prototypical cases resulting from sense extensions (e.g. *cirque* 'mess', semantically extended by metaphor from *cirque* 'circus').

Importantly, such carefully curated data, when not supported by existing semantic resources, is time-consuming to produce and requires specific linguistic expertise. While developing a resource like SuperWikt-fr involves similar efforts for annotating training data, it results in a more generic, large-scale resource. Here, we demonstrate this by extending Tribout et al. (2014)'s approach to a larger set of nouns, allowing us to test Croft's other hypotheses on the semantics of word classes. More broadly, we aim to show that resources like SuperWikt-fr are promising for conducting linguistics research based on large empirical datasets, and for enabling quantitative typological studies across a broader range of languages, beyond languages with large scale lexical semantic resources.

3 Building a SuperWik for French

We now describe the supervised classifier designed to automatically supersense-annotate the lexical senses of the French Wiktionary.

3.1 Data

3.1.1 The French Wiktionary

Wiktionary is a free, collaborative, online multilingual dictionary project created by the Wikimedia Foundation, available for various languages. The coarse structure of wiktionaries is shared across languages: an **entry** corresponds to a lemma and part-of-speech, and groups a list of **senses**. Each sense includes, among other things, a definition and **exemplars**, namely sentences illustrating the use of that sense in context. Several entries for the same lemma and with same part-of-speech correspond to homonymy, while several senses of an entry correspond to polysemy.

Although developed by non-specialists, we chose this resource for its free, comprehensive nature and for its overall quality. For French, in particular, the granularity and consistency of the sense inventory have been judged satisfactory, based on inter-annotator agreement in a verb disambiguation task in context (Segonne et al., 2019). We used the Dbnary turtle format (Sérasset, 2012) of the French Wiktionary, restricting our work to common nouns[3].

3.1.2 Supersense inventory for French nouns

We use the semantic class inventory defined by Barque et al. (2020), comprising 24 simple supersenses (grouped into 9 hypersenses), adapted from WordNet's Unique Beginners. Based on their frequency in the FrSemCor corpus annotated by these authors, we ignored the Tops supersense, and merged Group and Part with Quantity. For the purposes of our study (section 5), we will use a third level of generalization with Croft's tri-partite classification into Object/Action/Property. The complete hierarchy of super-, hyper-, Croft classes used in our paper is presented in Table 1. Barque et al. (2020) have also defined complex types that can be created productively. We have retained only the three complex types identified as frequent in FrSemCor (Act*Cognition, Artifact*Cognition and Group×Person)[4], leading to a total of 24 supersenses either simple or complex.

Supersense	Hypersense	Croft's class
Animal	Animate	Object
Person		
Artifact	Inanimate	
Body		
Food		
Object		
Plant		
Substance		
Cognition	Information	
Communication		
Act	Dynamic_sit.	Action
Event		
Phenomenon		
Attribute	Stative_sit.	Property
Feeling		
Relation		
State		
Quantity	Quantity	-
Institution	Institution	
Possession	Possession	
Time	Time	

Table 1: Super- and Hyper-sense inventories we use, along with the grouping into Croft's classes

3.1.3 Manually supersense-annotated senses

In order to train and evaluate a classifier of senses into supersenses, we manually annotated a set of FrWiktionary nominal senses. To limit the annotation effort, we started from a set of noun senses annotated with hypersenses by Aloui et al. (2020), and further annotated them with supersenses, based on the sense definition and its exemplars if any. Because Aloui et al. (2020) selected nouns based on their frequency but also favoring those having senses of a single hypersense, the distribution of supersenses in this data is not representative of the lexicon. We thus annotated two additional sets of senses: a **Random set** of senses, randomly selected from the entire FrWiktionary, and a **Frequency set**, for which we first randomly selected nouns from a list of 10000 most frequent French nouns[5]. Evaluation on the former set is meant to provide an approximation of the quality over the full lexicon, while evaluation on the latter focuses on more frequent

[3] In all the following, we report work using the 2023-03-20 dump for French, available at https://kaiko.getalp.org/about-dbnary/download/.

[4] The corpus contains annotations for 64 complex supersenses, most of them occurring once.

[5] The frequencies were calculated in an extract from Wikipedia, Wikisource and Uncorpus (three extracts of 53, 300 and 134 million tokens respectively), in the French part of the BigScience corpus, available at https://huggingface.co/spaces/bigscience-data/bigscience-corpus.

senses. The manual annotation was performed by one of the authors. To evaluate the reliability of the annotations, a random sample of 41 nouns, corresponding to 204 senses, was co-annotated with another lexical semantics expert in a double-blind procedure, resulting in a raw agreement of .76 and a Cohen's kappa of .74.

Set	Lem.	Senses		Ex.
		total	w. ex.	
Train	4,012	10,117	8,438	20,265
Freq-dev	465	1,581	1,365	3,638
Freq-test	448	1,339	1,154	3,026
Rand-dev	472	540	278	491
Rand-test	473	649	365	630
Total	5,870	14,226		

Table 2: Statistics in the supersense-annotated sets: number of noun lemmas (Lem.), number of senses in total, and having at least one exemplar (w. ex.), and total number of exemplars (Ex.)

We then designed a split of all these supersense-annotated senses into training set, frequency development and test sets, and random development and test sets, ensuring that these five sense sets correspond to disjoint sets of nouns. The statistics are presented in Table 2. Overall, the random and frequency sets have quite different distributions of supersenses (as shown in Appendix, in Table 10). In particular, the high proportion of Persons in the random dev and test sets (34% and 28%) reflects a massive presence of demonyms in FrWiktionary.

Moreover, the average number of senses per noun varies across the training, frequency, and random sets (respectively having a ratio of 2.5, 3.2 and 1.26 senses per lemma). The well-known higher polysemy for frequent words is confirmed. The ratio for the union of Rand-dev and Rand-test (1.26) is close to that computed on the full SuperWikt-fr (1.31, as shown later on in table 6, section 4). The polysemy level of the training set is in-between frequent nouns and random nouns.

3.2 Supervised classification

Our goal was to develop a system that takes a sense as input and returns one of the 24 selected supersenses. We explored different supervised classifier variants, differing mainly in the type of input: a definition versus an exemplar. These two input types are linguistically quite distinct. The definition describes the sense but generally does not include the word itself, whereas the exemplar contains the word, inflected, in context. As a result, we decided to train separate classifiers: we will refer to these as **`def`** versus **`ex`** classifiers. In the former case, we also compared giving only the definition as input (**`def`**), or the concatenation of the lemma, ':' and the definition (hereafter **`lem+def`**). Note the **`ex`** variant only applies to senses having at least one exemplar, and predicts the supersense for a given sense s, by averaging the log-probability scores over all the exemplars of s.

The architecture is similar for the three variants: the input sequence is passed into a French specific pretrained bidirectional language model (FlauBERT (Le et al., 2020), flaubert_large_cased), then the embedding for the relevant token, at the last layer, is passed into a multi-layer perceptron (MLP) with a single hidden layer. For definitions, the selected token is the beginning-of-sequence special token, whereas for exemplars, the selected token is the first subword token from the first occurrence of the relevant word in the text.

3.2.1 Experiments for each classifier variant

We trained the classifiers using a NLL loss, with non-frozen FlauBERT's parameters, and a dropout layer after the hidden layer.

We used a grid search to tune the hyperparameters[6], launching 10 runs for each combination, for each of the three variants. Table 3 shows the highest and average accuracies on the development sets, using the optimal hyperparameter combination.

Accuracies are higher on the random set than for the frequency set. Differences in performance were to expect, given the two sets have different supersense distributions (see Table 10 in Appendix). The random set contains much more animate entities, and these reveal much easier to classify (cf. F-scores above 96% for animate entities in the performance break down, Table 11, in the Appendix, for our best classifier). On the contrary, the frequency set contains more Cognition senses, which are classified with less accuracy.

Among the three classifier variants, using definitions as input significantly outperforms using exemplars, with the best run on freq-dev achieving 73.1% for `def` compared to 65.0% for `ex`. This demonstrates that the highly constrained form of definitional paraphrases makes semantic classifica-

[6] We tested learning rate=[5e-5, 1e-5, **5e-6**, 1e-6], hidden layer size=[256, 512, **768**], dropout rate=[0.1, **0.3**]. The best combination (in bold) is the same for the three variants.

	Rand-dev			Freq-dev		
	Mean (MAD)	Best run		Mean (MAD)	Best run	
	Super	Super	Hyper	Super	Super	Hyper
`ex`	63.9 (1.2)	65.6	77.4	64.2 (0.7)	65.1	72.5
`def`	78.3 (0.7)	80.0	86.5	71.3 (1.1)	73.1	78.9
`lem+def`	**82.1** (0.7)	**83.3**	**90.6**	**74.8** (0.8)	**76.7**	**82.2**

Table 3: Accuracies in % (mean and MAD on ten runs, and best run results), on the Random and Frequency development sets for both supersense (Sup.) and hypersense (Hyp.) granularity, using the lexicographic exemplars alone (**`ex`**), the definition alone (**`def`**), and the lemma concatenated with the definition (**`lem+def`**).

tion easier compared to target words used in context. Concatenating the lemma to the definition (`lem+def`) further increases performance, for both random and frequency sets, with gains of 3 to 4 points. This could be due to semantic generalization learnt in the pre-trained language model.

In order to evaluate the necessary amount of training senses needed to reach high performance, we provide a learning curve in Figure 1 in the Appendix, for the `lem+def` variant. It suggests that while more annotated senses could further increase performance a little, 4000 examples are sufficient to achieve a performance rather close to that achieved with all training data.

3.2.2 Combining architectures

The next step is to select the best classifier (the best run) to annotate the nouns of the entire Fr-Wiktionary. Given the results, we selected the best `lem+def` run, and tried to further improve it by combining it with the best `ex` run, as these two variants are likely to be complementary. We thus tested a **`lem+def&ex`** variant, which scores each supersense by weighting the log-probability scores of `lem+def` and `ex`, using their respective accuracies as weights[7]. We provide the performance obtained on dev sets in the first row of Table 4.

Given that the performance is indeed better, this is the classifier we used to annotate the entire resource. Results on the evaluation test sets (last row of Table 4) provide the best approximation of the quality of supersense annotations on the full resource[8]. Note though performance varies across supersenses, animate entities being the easiest senses to classify, followed by inanimate entities (as shown in the break-down per supersense, Table 11, in Appendix).

	Rand		Freq	
	Sup.	Hyp.	Sup.	Hyp.
dev sets	84.3	91.3	77.1	83.0
test sets	84.8	91.9	80.3	86.2

Table 4: Accuracies in %, for both supersense (Sup.) and hypersense (Hyp.) granularity, on the Random and Frequency development and test sets, using the classifier retained to tag the whole resource (**`lem+def&ex`**).

3.2.3 Analysis

Table 5 breaks down the performance of the combined best model (`lem+def&ex`) based on whether the lemmas of the classified senses are simple or polylexical (MWEs), and whether they are monosemic or polysemous.

At first glance, MWEs (present only in the random set) seem more difficult to classify. A closer examination reveals that this is rather due to a different supersense distribution for simple versus MWE lemmas: senses from easier-to-classify supersenses, such as Person, are far more numerous among monolexical lemmas, boosting performance in this set.

Moreover, performance for the senses of monosemic lemmas is higher than for those of polysemous lemmas, as shown in the bottom part of Table 5 (+6 pt on average). The static representation of the lemma concatenated with the definition,

[7] Namely 83.3 and 65.7, divided by their sum. We defaulted to the `lem+def` score for senses having no exemplar.

[8] The difference between IAA (76%, cf. Section 3.1.3) and accuracy on Freq-test (80.3%) can be explained by the fact that the tagset used for manual annotation reliability assessment was open to complex supersenses. Among the 204 double-blind annotated senses, annotator#1 annotated 23 senses with 9 different complex supersenses, while annotator#2 used 6 complex supersenses across 20 senses. As expected, complex supersenses were involved in a large proportion of disagreements (20/49). In contrast, only the three most frequent complex supersenses (act*cognition, artifact*cognition, groupxperson) were retained for the experiments (Section 3.1.2), thus making the task slightly easier.

which aggregates the different senses if the form is ambiguous, may explain this discrepancy.

	Rand-dev		Freq-dev	
	Sup.	Hyp.	Sup.	Hyp.
All	84.3	91.3	77.1	83.0
Simple lemma	85.5	92.3	77.1	83.0
MWE	77.4	85.7	-	-
1-sense lemma	85.6	91.9	82.5	87.5
2+-sense lemma	77.1	87.9	76.1	82.1

Table 5: Accuracy (%) of the best model (def+lemma&ex) on the Random and Frequency dev sets, for both supersenses (Sup.) and hypersenses (Hyp.), depending on whether the lemmas of the classified senses are simple lemmas or multi-word-expressions (MWEs), and whether they are monosemic or polysemous.

4 The SuperWikt-fr

We applied our best classifier to supersense-tag all the nominal senses of the FrWiktionary, and we now provide a few statistics concerning the resulting resource, which we refer to as "SuperWikt-fr". As presented in Table 6, around 306k nominal senses were tagged, corresponding to about 229k nouns (nominal lemma types). Strikingly, a massive proportion (20%) of these senses correspond to the particular class of demonyms[9]. Homonymy (nouns with several entries) concerns only 2% of the nouns, and so is quite rare at the level of the whole lexicon. Ambiguity is much more present, concerning 16% of the nouns, these having an average ambiguity of 3.1 senses per noun. Overall the average ambiguity is 1.34. These statistics indicate that the size and the sense granularity is roughly comparable to the English WordNet 3.0 (149k nouns, 1.24 noun-synset pairs on average[10]).

We provide the percentages of annotated supersenses and hypersenses in Tables 7 and 8 respectively. The very high percentage of the Person supersense reflects the massive presence of demonyms. The next two prominent senses are Artifact and Act.

Nb of nouns	229,174
Nb of nominal entries	234,172
Nb of senses	306,530
Avg nb of senses per noun	1.34
Avg nb of senses per entry	1.31
Homonymous nouns (>1 entry)	2%
Ambiguous nouns (>1 sense)	16%
Polysemous entries (>1 sense)	15%
Nominal MWE	21%
Senses without exemplars	51%
Demonym senses	20%

Table 6: Statistics of the supersense-annotated FrWiktionary, for nouns. MWE are identified as lemmas containing a space and/or a quote.

Supersense	%	Example
Act	9.26	*contrôle* ‘control’
Act*Cognition	0.44	*discours* ‘speech’
Animal	6.83	*mouton* ‘sheep’
Artifact	12.93	*chapeau* ‘hat’
Artifact*Cognition	0.75	*livre* ‘book’
Attribute	2.53	*taille* ‘size’
Body	2.02	*rein* ‘kidney’
Cognition	6.24	*idée* ‘idea’
Communication	2.55	*braille* ‘Braille’
Event	2.13	*famine* ‘famine’
Feeling	0.43	*joie* ‘joy’
Food	2.45	*pain* ‘bread’
Group×Person	0.34	*foule* ‘crowd’
Institution	1.83	*banque* ‘bank’
Object	2.79	*ruisseau* ‘brook’
Person	31.99	*mère* ‘mother’
Phenomenon	0.62	*lueur* ‘glow’
Plant	3.64	*chêne* ‘oak’
Possession	0.91	*taxe* ‘tax’
Quantity	1.30	*tonne* ‘tonne’
Relation	0.21	*rapport* ‘link’
State	2.51	*solitude* ‘loneliness’
Substance	4.37	*colle* ‘glue’
Time	0.82	*printemps* ‘spring’

Table 7: Percentages of predicted supersenses in SuperWikt-fr.

5 Use case: assessing Croft’s hypothesis

We will now leverage SuperWikt-fr to both test and quantify the three predictions drawn from (Croft, 1991) (cf. section 2), for French nouns. The first, already evaluated on a smaller scale in (Tribout et al., 2014), is that simplex nouns essentially fall

[9] After analysis of the definition of a few demonyms, we concluded we could evaluate this proportion simply by counting senses whose definition starts with *[Hh]abitant** ‘inhabitant’, thus matching both masculine and feminine forms.

[10] https://wordnet.princeton.edu/documentation/wnstats7wn

Hypersense	Proportion
Animate entity	38.82
Inanimate entity	28.21
Dynamic situation	12.01
Informational object	8.79
Stative situation	5.68
Institution	1.83
Quantification	1.30

Table 8: Percentages of hypersenses in SuperWikt-fr, for the hypersenses covering more than 1% of the senses.

into the semantic class of objects. The other two are that nouns with an action (resp. property) sense are prototypically derived from verbal (resp. adjectival) bases. Note that Croft (1991) does not provide explicit definitions for the three semantic classes he mentions. The Object class is understood here to encompass both animate and inanimate entities, as well as informational objects (e.g., *idée* 'idea'). The Action class includes all nouns that denote dynamic situations, whether agentive (e.g., *bombardement* 'bombing') or not (e.g., *storm* 'orage'). The Property class comprises nouns that refer to stative situations, whether transitional (e.g., *solitude* 'loneliness') or not (e.g., *intelligence* 'intelligence'), cf Table 1.

5.1 Data selection

To conduct our study, we need morphological information on words, which is not available in SuperWikt-fr. We use data from two recently released French resources on derivational morphology, namely *Demonette-2* (Namer et al., 2023) and *Échantinom* (Bonami and Tribout, 2021). These datasets specify whether a noun is simplex or morphologically complex, and in the latter case, the type of morphological construction from a base, and the part-of-speech of this base. Furthermore, the nouns considered in the study must be sampled to approximate the distribution of words according to their frequency of use, which is also not the case with SuperWikt-fr. We have therefore chosen to focus on the nouns in the *Lexique-3* lexicon (New et al., 2007), which lists all nouns appearing in large corpora of contemporary French (literary corpora and/or film subtitle corpora).

The intersection of nouns from the wiktionary (229,174), Lexique-3 (30,567) and the combination of the two morphological resources (53,033 ∪ 5,000) provides an initial list of 17,474 nouns. We then focused exclusively on the two most frequent morphological classes—simplex and suffixed nouns—which together account for 83.1% of the total[11]. In the resulting set, we further discarded senses having a complex supersense, or a too infrequent hypersense. The final subset (hereafter the **study subset**) comprises 13,945 nouns, corresponding to 34,829 senses in SuperWikt-fr. An assessment of the semantic classification quality for this set can be obtained by averaging the hypersense F-scores obtained by our classifier, each weighted by its frequency, which gives an F-score of 86.8 for hypersenses.

5.2 Data analysis

Table 9 provides a breakdown of the data according to their morphological and semantic classes (Croft's classes). Regarding morphological classes, we observe that nominal senses are more associated with suffixed lemmas than with simplex ones (cf. last column of Table 9, 59.1% versus 41.9%). Semantically, most of nominal senses fall under the Object class (63.3%), followed by Action (24.5%) and Property (12.1%).

5.2.1 Semantic properties of simplex nouns

Let us first examine the distribution of semantic classes among simplex nouns (see the percentages in round brackets in the first line of Table 9). We observe that 80.7% denote an object, 11.5% denote an action, and 7.8% denote a property. These results thus totally support Croft's hypothesis that simplex nouns typically denote objects.

Focusing on Action and Property senses, atypical among simplex nouns, one can further wonder whether or not they are extended from another sense of the lemma (hereafter, whether the sense is **extended** or **primary**). For example, the Action sense of *crime* 'crime' is primary, while the actional sense of *pont* 'long weekend' is extended by metaphor from the object sense *pont* 'bridge'. Similarly, the Property sense of *grâce* 'grace' is primary, while *robe* 'wine or horse color' is derived from the object sense *robe* 'dress'. For a first rough estimation of how many atypical senses of simplex nouns are in fact extended senses, we can use the hypothesis that a sense listed first in the SuperWikt-fr entry

[11] We excluded cases of conversion from another part-of-speech (14.2%), for which defining the direction of the derivation is tricky (e.g Balteiro, 2007), as well as nouns formed through other, less common morphological processes such as prefixation and compounding (2.7%).

	Action			Object			Property			Total
Simplex	4.7	[19.1]	(11.5)	33.0	[52.2]	(80.7)	3.2	[26.4]	(7.8)	**40.9** (100)
Suffixed	19.9	[80.9]	(33.6)	30.3	[47.8]	(51.2)	8.9	[73.6]	(15.1)	**59.1** (100)
Total	**24.5**	[100]		**63.3**	[100]		**12.1**	[100]		**100.0**

Table 9: Percentages of semantic and morphological classes, for all the senses (34,829) in the **study subset**. Round brackets (resp. square brackets) provide proportions within columns (resp. rows). Example of reading: the upper left cell "4.7 [19.1] (11.5)" means that 4.7% of senses were annotated with the Action class, and are senses of a simplex noun, which corresponds to 19.1% of Action senses, and to 11.5% of senses of simplex nouns.

is primary[12], while the other senses in the same entry are obtained by sense extension (but not necessarily from the first sense). We obtain that among action senses of simplex nouns, only 35.6% are listed first in their entry. For the 64.4% supposed to be extended action senses, the sense source of the extension is difficult to determine precisely, but as a first approximation, we observe that 36.1% of them are in an entry whose first sense is an Object (e.g., *pont* ‘bridge/long weekend’). Among these, we find cases of the regular metonymy producing an activity sense from an artifact sense (e.g. *aviron* ‘rowing’ sense derived from ‘oar’ sense, *piscine* ‘swimming’ sense derived from ‘swimming-pool’). As for senses of simplex nouns denoting Property, 36.8% are primary, whereas 63.2% are extended. Among the latter, 33.8% are senses of lemmas having an object primary sense (e.g., *robe* ‘dress/color of a wine’).

Even though we use automatically annotated semantic classes and an indirect estimation of sense extension, we believe that thanks to the scale of our study, these observations concerning the origin of atypical non-object simplex senses reinforces Croft’s hypothesis. Not only are these atypical senses very much in the minority within the class of simplex nouns, but a significant proportion of them (between a quarter and a third) are likely to result from a sense extension.

5.2.2 Morphological properties of nouns having non-objectal senses

As a reminder, according to Croft’s predictions, action senses are prototypical of the verb category. Therefore, action nouns are expected to be predominantly derived from verbs. The data in Table 9 (first column, percentages in square brackets) confirms that action senses are overwhelmingly associated with suffixed nouns (80.9%) and far less frequently with simplex nouns (19.1%). Furthermore, within the former group, 88.2% of action senses are linked to lemmas derived from verbs, thus confirming and quantifying Croft’s predictions.

Similarly, Property senses being not prototypical of nouns, Croft predicts that they should be mostly derived from adjectives. Here again, the data presented in Table 9 (third column, percentages in square brackets) confirms and quantifies this prediction: property nominal senses are much more senses of suffixed nouns (73.6%) than of simplex nouns (26.4%). However, the morphological origin of property senses associated to suffixed nouns is more contrasted: while a majority are associated with nouns derived from adjectives (53.6%), 28.8% are associated with nouns deriving from verbs and 11.9% from nouns. Property senses derived from verbs are mostly constructed with the suffix *-ion*, such as *exaltation* ‘elation’ from *exalter* ‘to elate’. Property senses associated with nouns suffixed from nouns are mostly built from the suffix *-ism*, such as *égoïsme* ‘selfishness’.

6 Conclusion

In this paper, we have demonstrated that a lexicon can be semi-automatically enriched with semantic annotations of sufficient quality to empirically test lexical semantics hypotheses. This approach can be applied to any language with access to a large scale dictionary, such as Wiktionary, and requires a

[12] In lexicographic practice, senses are typically ordered based on historical criteria (chronological attestation) or usage frequency (e.g. *WordNet*). French Wiktionary guidelines highlight the importance of these factors for sense ordering (https://fr.wiktionary.org/wiki/Convention:DÃĺfinitions). To provide a more objective assessment, we evaluated this approximation on the 41 randomly selected nouns used to calculate inter-annotator agreement (IAA), of which 11 are monosemous (entries with a single sense). For the 30 polysemous nouns, we determine if the first listed sense is semantically primary by assessing whether the other senses can be derived from it—either directly or indirectly—through processes like broadening, narrowing, metaphorical, or metonymic extension. The results show that the first listed sense is primary in 83.3% of cases, with the remaining cases being debatable.

reasonable amount of manual semantic annotation (in comparison to the annotation effort of WordNets: we showed that training on 4000 examples reaches an accuracy of more than 80%, for mono-label classification, with a tagset of 24 supersenses, training on the full set of 16000 examples reaches almost 85%).

The analyses conducted on a substantial subset of French nouns made it possible to check and quantify Croft's predictions that simplex French nouns mostly denote objects, and that French nouns denoting an action (respectively a property) are more frequently derived from verbs (respectively adjectives). More importantly, they highlighted the complex nature of the relationships between forms and meanings in the lexicon, emphasizing the role of polysemy and derivational morphology in meaning construction. Note that even though our observations are based on automatic semantic annotations, their evaluated quality (almost 85% for supersenses and almost 92% for hypersenses, on the random test set, table 4) allows to make large-scale observations. Moreover, we believe that the generic nature of the semantic annotations in SuperWikt-fr can serve for various other studies in lexical semantics, for instance providing access to nouns that follow regular polysemy patterns (e.g the Animal>Person metaphor), or to nouns having hybrid meanings (indicated by a complex supersense in the resource).

The resource could be enriched in many ways, in particular with frequency information, which is an important gap for empirical linguistic studies (both experimental and computational). SuperWikt-fr can be leveraged to learn a super- and hypersense tagger, by corpus-projecting the labels of monosemic nouns, in the manner of (Aloui et al., 2020). SuperWikt-fr's wide coverage allows us to hope for good tagging quality. The predicted hyper- and super-senses then enable partial sense disambiguation, which can be used to enrich SuperWikt-fr, or for WSD.

7 Limitations

The approach described in this work supposes the availability of a large-scale electronic dictionary, and is hence inapplicable for very low-resourced languages. Moreover, our work was restricted to nominal senses.

Our approach is supervised, and required us to manually tag word senses with supersenses. We measured that training on 4000 examples reaches an accuracy of more than 80%, using a tagset of 24 supersenses, and training on the full set of 16000 examples reaches almost 85%. Hence there is still large room for improvement of the quality of the annotated Wiktionary.

A final limitation concerns the method used to select the data in the use case (section 5), an issue recently brought to our attention by Olivier Bonami. The morphological information we relied on comes, for one third, from a resource that samples nouns from a large French corpus (Échantinom), and for two thirds, from a resource constructed using several existing morphological databases (Demonext). This approach carries the risk of underrepresenting simplex nouns in our dataset and, more broadly, of the data not being fully representative of the natural distribution of French nouns. Specifically, we observed a larger proportion of suffixed nouns compared to simplex nouns (59% vs. 41%) in our dataset, whereas the same ratio in Échantinom is closer to 51% vs. 49%[13]. We are currently working on improved data sampling to develop a morpho-semantic lexicon that better reflects the actual distribution of nouns in French, so that links between morphological information and supersenses provided in SuperWikt-fr can be more accurately quantified.

Acknowledgments

We are grateful to the anonymous COLING reviewers and to Olivier Bonami for their valuable comments on an earlier version of this study. Any remaining shortcomings are ours.
This work has been partially funded by public grants overseen by the French National Reseach Agency (ANR), both within the SELEXINI ANR project (ANR-21-CE23-0033) and the Empirical Foundations of Linguistics Labex (as part of the program "Investissements d'Avenir" (ANR-10-LABX-0083), which contributes to the IdEx Université Paris Cité (ANR-18-IDEX-0001)).

References

Cindy Aloui, Carlos Ramisch, Alexis Nasr, and Lucie Barque. 2020. SLICE: Supersense-based lightweight

[13] This ratio was calculated from the 1,531 suffixed nouns and 1,462 simplex nouns in Échantinom that are not described as having a morphological conversion relationship with another lexeme.

interpretable contextual embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3357–3370, Barcelona, Spain (Online).

I. Balteiro. 2007. *The directionality of conversion in English: A dia-synchronic study*. Peter Lang, Berlin.

Lucie Barque, Pauline Haas, Richard Huyghe, Delphine Tribout, Marie Candito, Benoit Crabbé, and Vincent Segonne. 2020. FrSemCor: Annotating a French corpus with supersenses. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5912–5918, Marseille, France.

O. Bonami and D. Tribout. 2021. Échantinom: A hand-annotated morphological lexicon of French nouns. In *International Workshop on Resources and Tools for Derivational Morphology*, pages 42–51, Nancy, France.

P. Buitelaar. 1998. Corelex: An ontology of systematic polysemous classes. In N. Guarino, editor, *Formal Ontology in Information Systems*, pages 221–235. IOS Press.

M. Ciaramita and Y. Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 594–602, Sydney, Australia.

M. Ciaramita and M. Johnson. 2003. Supersense tagging of unknown nouns in wordnet. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, pages 168–175.

W. Croft. 1991. *Syntactic categories and grammatical relations: the cognitive organization of information.* University Press of Chicago.

S. Dei Rossi, G. Di Pietro, and M. Simi. 2011. Evalita 2011: Description and results of the supersense tagging task.

D. Fiser and B. Sagot. 2015. Constructing a poor man's WordNet in a resource-rich world. *Language Resources and Evaluation*, 49(3):601–635.

L. Flekova and I. Gurevych. 2016. Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2029–2041, Berlin, Germany.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, pages 2479–2490, Marseille, France.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online.

G. Miller, R. Beckwith, C. Fellbaum, Gross D., and K. Miller. 1990. Wordnet: An online lexical database. *International Journal of Lexicography*, (3):235–244.

V. Mititelu, S. Leseva, and I. Stoyanova. 2021. Semantic analysis of verb-noun derivation in princeton wordnet. In *Proceedings of the 11th Global Wordnet Conference*, pages 108–117, University of South Africa, South Africa.

F. Namer, N. Hathout, D. Amiot, L. Barque, O Bonami, G. Boyé, B. Calderone, J. Cattini, G. Dal, A. Delaporte, G. Duboisdindien, N. Falaise, A. Grabar, P. Haas, F. Henry, M. Huguin, N. Juniarta, L. Liégeois, S. Lignon, L. Macchi, G. Manucharian, C. Masson, F. Montermini, N. Okinina, F. Sajous, D Sanacore, T. M. Tran, J. Thuilier, Y. Toussaint, and D. Tribout. 2023. Démonette-2, a derivational database for french with broad lexical coverage and fine-grained morphological descriptions. *Lexique*, 33:6–40.

B. New, M. Brysbaert, J. Veronis, and C. Pallier. 2007. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28(4):661–677.

B. S. Pedersen, A. Braasch, A. Johannsen, H. Martinez Alonso, S. Nimb, S. Olsen, A. Søgaard, and N. Hartvig Sørensen. 2016. The semdax corpus – sense annotations with scalable sense inventories. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 842–847, Portorož, Slovenia.

W. Peters and I. Peters. 2000. Lexicalised systematic polysemy in WordNet. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559, San Diego, California.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A. Smith. 2012. Coarse lexical semantic annotation with supersenses: An Arabic case study. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 253–258, Jeju Island, Korea.

V. Segonne, M. Candito, and B. Crabbé. 2019. Using Wiktionary as a resource for WSD : the case of French verbs. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 259–270, Gothenburg, Sweden.

G. Sérasset. 2012. Dbnary: Wiktionary as a LMF based multilingual RDF network. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2466–2472, Istanbul, Turkey.

D. Tribout, L. Barque, P. Haas, and R. Huyghe. 2014. De la simplicité en morphologie. In *SHS web of conferences*, volume 8, pages 1879–1890. EDP Sciences.

P. Vossen. 1998. Introduction to EuroWordNet. In *EuroWordNet: A multilingual database with lexical semantic networks*, pages 1–17. Dordrecht: Kluwer.

A Appendix

	Train	Dev		Test	
		Freq	Rand	Freq	Rand
Supersense					
Act	11.1	19.5	7.6	23.0	8.3
Animal	3.9	0.8	5.7	0.8	4.0
Artifact	12.4	13.9	9.6	16.1	19.4
Attribute	6.7	9.4	3.0	7.9	2.9
Cognition	9.2	11.6	6.9	11.4	6.0
Person	9.5	7.5	34.4	6.6	27.6
State	4.2	4.8	3.1	5.5	2.0

Table 10: Distribution of supersenses in the five manually annotated sets (keeping only supersenses appearing more than 5% in any set).

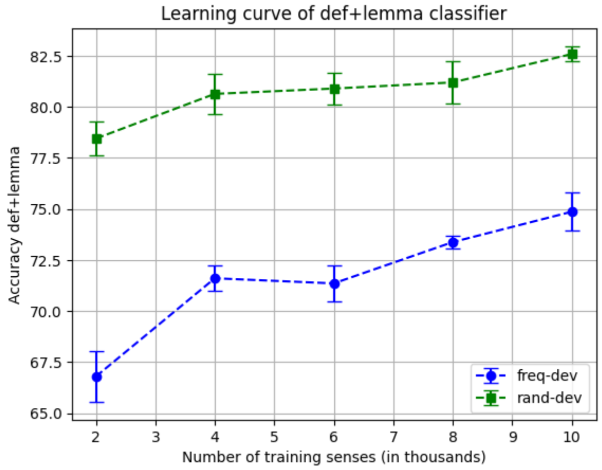

Figure 1: Learning curve for the `lem+def` variant (average over 5 runs, accuracy of supersense prediction, on the Random and Frequency development sets).

Hypersense	**Supersense**	**rand dev**		**freq dev**	
Animate entity	Animal	97.7	90.9	96.6	100.0
	Person		98.4		96.2
Inanimate entity	Artifact	93.9	88.9	90.9	86.3
	Body		76.9		84.9
	Food		76.2		85.7
	Object		57.7		68.4
	Plant		75.0		96 .5
	Substance		75.0		81.4
Dynamic situation	Act	89.2	87.1	86.7	85.9
	Event		75.7		70.0
	Phenomenon		0.0		48.3
Stative situation	Attribute	78.1	83.9	79.7	70.4
	Feeling		85.7		64.0
	Relation		NA		29.6
	State		53.8		62.2
Informational object	Cognition	77.5	66.7	69.9	65.8
	Communication		69.2		74.4
Quantification	Quantity	85.7	85.7	61.2	61.2
Institution	Institution	57.1	57.1	68.1	68.1
Possession	Possession	71.4	71.4	81.8	81.8
Time	Time	66.7	66.7	72.2	72.2
Dynamic situation*Informational object	Act*Cognition	66.7	66.7	53.1	53.1
Inanimate entity*Informational object	Artifact*Cognition	100.0	100.0	73.7	73.7
Quantification×Animate entity	Group×Person	0.0	0.0	84.7	84.7

Table 11: F-scores for each supersense and hypersense (in %), for the random and frequency dev sets, with the best classifier lem+def&ex.