Marine Wauquier, Nabil Hathout & Cécile Fabre

Semantic discrimination of technicality in French nominalizations

Abstract: French suffixations in *-age*, *-ion* and *-ment* are considered roughly equivalent, yet some differences have been pointed out regarding the semantics of the resulting nominalizations. In this study, we confirm the existence of a semantic distinction between them on the basis of a large scale distributional analysis. We show that the distinction is partially determined by the degree of technicality of the denoted action: *-age* nominals tend to be more technical than *-ion* ones. We examine this hypothesis through the statistical modeling of technicality. To this end, we propose a linguistic definition of technicality, which we implement using empirical, quantitative criteria estimated in corpora and lexical resources. We show to what extent the differences with respect to these criteria adequately approximate technicality. Our study indicates that this definition of technicality, while amendable, provides new perspectives for the characterization of action nouns.

Keywords: action nouns, derivation, technicality, distributional semantics, statistical modeling

1. Introduction

Among the various French derivational processes available to coin deverbal action nouns, the suffixations in *-age* (*remplir* 'to fill' → *remplissage* 'filling'), *-ion* (*réduire* 'to reduce' → *réduction* 'reduction') and *-ment* (*allonger* 'to lengthen' → *allongement* 'lengthening') are the most productive, and are often said to be rival. Several semantic differences have been put forward to explain the suffix rivalry between them.

Some works point towards the nature of the base verb arguments. Kelling (2001) and Martin (2010) claim that verbs involved in *-age*, *-ion* and *-ment* suffixations differ in the agentivity degree of their subjects, while Fradin (2014) argues for a difference of concreteness of the referent denoted by their objects. Other works focus on the semantic properties of the base verb or on the nominalizations themselves. Dubois (1962) and Martin (2010) discuss the preference of the suffix *-age* for verbs that denote physical actions, and the use of *-age* nominalizations for industrial processes. They also acknowledge that the suffix *-ion* is more frequent in the scientific terminology. As for the suffix *-ment*, it is said to be unmarked at the ontological level – i.e. it does not select a domain-specific reading as *-age* does – (Martin 2010), and its nominalizations tend to denote attitude or psychological states (Dubois 1962).

 https://doi.org/10.3726/zwjw.2020.02.06

In the light of the diachronic evolution of the suffix *-age*, Fleischman (1990), cited by Uth (2010), makes a claim similar to Dubois's that *-age* nominalizations are strongly associated to the industry domain. Fleischman states that the number of *-age* nominalizations increased in the 19th century, brought about by the industrial revolution and the growing need to designate new technologies and technical processes – which was later demonstrated by Uth (2010). According to Fleischman (1990), French borrowed part of its terminology from English, which extensively used the *-age* suffix, previously borrowed from French. Although the borrowing hypothesis has not been proved, the suffix *-age* is still very productive in French. This raises a question: do lexicalized and neological *-age* nominalizations still tend to be more technical than *-ion* and *-ment* nominalizations?

In this study, we investigate the specialization of derived action nouns in terms of technicality, and more specifically the hypothesis that *-age* action nouns have a higher degree of technicality. We show that this notion plays a significant role in the distinction between nominalizations in *-age*, *-ion* and *-ment*.

We first use a distributional semantic model of a contemporary French corpus to corroborate the existence of a difference in the degree of technicality of French nominals (Section 2). We then propose a definition of technicality and infer a set of linguistic corollaries that we translate into empirical criteria evaluated from corpora and lexical resources (Section 3). Finally, we assess the capacity of these criteria to gauge noun technicality and discriminate the three suffixes (Section 4). The results suggest that our definition of technicality can partially contribute to characterize *-age*, *-ion* and *-ment* deverbal action nouns insofar as *-age* nominalizations tend to be more technical than *-ion* and *-ment* nominalizations. This work is exploratory in nature and provides foundations for a new perspective on the semantic distinction between *-age*, *-ion* and *-ment*.

2. Distributional discrimination of action nouns

Our study of the notion of technicality and its linguistic expression stems from observations, made by means of distributional semantic, that confirm the existence of distributional differences between French nominalizations in *-age*, *-ion* and *-ment*. This section is dedicated to the presentation of these initial observations. The semantic distributional model we use was created with Word2Vec (Mikolov et al. 2013). Our experimental set-up also includes Lexeur, a lexical resource designed for the comparison of properties of words that belong to the same derivational family. We first describe these two resources before presenting the hypothesis and the first observations.

2.1. Lexeur

Our study is based on Lexeur, a French morphological database[1] which contains 5.974 derivational subfamilies of agent nouns suffixed in *-eur*.

Lexeur subfamilies are made up of: (i) an entry, namely a masculine noun in *-eur* (*abatteur* 'slaughterer', *camionneur* 'truck driver', *prédateur* 'predator'); (ii) one or several feminine equivalents (*abatteuse* 'female slaughterer' or 'harvester', *camionneuse* 'female truck driver', *prédatrice* 'female predator'); (iii) its verbal or nominal base if it exists (*abattre* 'to cut down' or 'to slaughter', *camion* 'truck'; *prédateur* does not have an attested base lexeme in French); (iv) a list of morphologically related verbs (*sélectionner* 'select' for *sélecteur* 'selector') and (v) a list of morphologically related nouns (*rectorat* 'board of education' for *recteur* 'superintendent of schools') when the entry is denominal or does not have an attested base lexeme; and (vi) a list of nominalizations of the base or nominals related to the entry (*abattage* 'slaughter', *prédation* 'predation'). All the fields except the first may be empty. The nouns in *-eur* were initially extracted from the *Frantext* corpus and from the French dictionary *Trésor de la Langue Française* (Dendien and Pierrel 2003). They were supplemented with words collected from the Web by annotators. All lexemes of the resource are associated with a morphosyntactic description in the Multext-Grace format. An excerpt of Lexeur is shown in Table 1.

Tab. 1: Five entries from Lexeur.[2]

MascAgt	sculpteur/ Ncms	inflammateur/ Ncms	inflammateur/ Ncms	autostoppeur/ Ncms	chiropracteur/ Ncms
FemAgt	sculpteuse/ Ncfs sculptrice/ Ncfs	inflammatrice/ Ncfs	inflammatrice/ Ncfs	autostoppeuse/ Ncfs	chiropractrice/ Ncfs
Base	sculpter/ Vmn----	enflammer/ Vmn----	inflammer/ Vmn----	autostop/ Ncms	
Nominals	sculpture/ Ncfs sculptage/ Ncms	inflammation/ Ncfs	inflammation/ Ncfs		chiropraxie/Ncfs

1 The resource was created in 2001 in the ERSS lab of CNRS (now CLLE lab) and the University of Toulouse Le Mirail (now University of Toulouse – Jean Jaurès) and is distributed under a creative commons license in a tabular separated format (tsv) and xml on the REDAC repository (http://redac.univ-tlse2.fr).

2 The fields Associated Verbs and Associated nouns are omitted because they are empty in all five entries. Masculine nouns are tagged Ncms, feminine nouns, Ncfs and infinitive verbs Vmn----.

Table 1 shows the diversity of the derivational families included in Lexeur: some columns are complete such as *sculpteur* 'sculptor', others are not, such as *chiropracteur* 'chiropractor' (with no identified base) or *auto-stoppeur* 'hitchhiker' (which has a nominal base but no other related nominal). 78% of the derivational subfamilies in the resource have a verb base, 14% are derived from a noun, and 8% are not associated with a base of any kind (such as *chiropracteur*). Some subfamilies include several nominals, such as *sculpteur* 'sculptor' (*sculpture*, *sculptage* 'sculpture'), while others don't have any, such as *auto-stoppeur* 'hitchhiker'. Moreover, some nouns in *-eur* have several bases and are part of several subfamilies (*inflammateur* 'igniter' may be derived from *inflammer* or *enflammer*, both equivalents of 'to ignite'). For these nouns, each base yields one derivational subfamily (i.e. one record).

2.2. Word2Vec

Distributional semantics is based on the hypothesis, known as the Distributional Hypothesis, that words that are semantically similar share similar distributions in corpora (Harris 1954; Firth 1957). Distributional Semantic Models (DSMs) represent words by vectors, called word embeddings, computed from all the contexts of those words in a corpus. Words with similar contexts are represented by close vectors and words with distinct contexts are represented by distant vectors. Being mathematical objects, word embeddings can be added, subtracted or multiplied in order to form analogies or compute the meaning of a phrase (Lenci 2018; Boleda 2020). We can also easily compute a distance between two vectors; the values range from 0 (no proximity) to 1 (strict equality).

The DSM we use in our study is built from the 2018 French edition Wikipedia. This corpus contains about 900 million words and has a large and diverse vocabulary that covers many domains and is in line with the vocabulary found in Lexeur. The corpus was lemmatized with Talismane (Urieli 2013). In order to limit the effects of the inherent instability of DSMs (Pierrejean 2020), we computed and concatenated 5 matrices into one unique DSM[3]. Each of these matrices is created by a new Word2vec run with all parameters set to their default values: CBOW architecture, Negative Sampling algorithm, a frequency threshold of 5, a window size of maximum 5, and 100 dimensions.

3 The DSMs were built using the OSIRIM computing platform that is administered by the IRIT computer science lab and supported by the National Center for Scientific Research (CNRS), the Région Midi-Pyrénées, the French Government, and the European Regional Development Fund (ERDF).

2.3. Methodology

Once the model is built, a proximity measure can be computed between each pair of words in the corpus. We can also identify the nearest neighbors of each word, i.e. the words that are the most distributionally similar (and the most semantically similar, according to the Distributional Hypothesis mentioned in Section 2.2.). For example, the nearest neighbors of *laminage* 'rolling' in our DSM are *forgeage* 'forging' (0.87), *formage* 'shaping' (0.85), *étirage* 'stretching' (0.81) and *extrusion* 'extrusion' (0.81), all denoting other steel or materials deformation processes.

The DSM can also be used to study distributional similarity at the level of noun classes by summarizing the information that is available at the level of the individual words. The idea is to represent each class by a prototypical representation of all the nouns in *-age*, *-ion* and *-ment*. If a semantic distinction exists between the three suffixations, as mentioned in Section 1, it should be reflected in the nouns coined by each suffixation, and be visible in the DSM. More specifically, a class of nouns may be represented by the average vector of the vectors of all its members (Kintsch 2001). The semantic content associated with this average vector is then characterized by its nearest neighbors (Marelli and Baroni 2015), since they are the words that are the most similar to the prototype of the class, and can therefore be considered to be representative instances of the class.

To construct these prototypical representations of *-age*, *-ion* and *-ment* nominalizations, we first need to select the members of each class. The nouns are extracted from Lexeur in the 'nominals' field of subfamilies with verb bases. At this stage, we have 4266 distinct deverbal nominals, including respectively 1687 nominalizations suffixed in *-age*, 1357 in *-ion* and 1222 in *-ment*. We decided to make a preliminary filtering before computing the *-age*, *-ion* and *-ment* centroids in order to build a more homogeneous distributional class for each suffix. We performed some manual filtering to exclude the nouns that: (i) are weakly related to the base verb indicated in Lexeur (*reportage* 'report' with respect to *reporter* 'to postpone'; 'to transfer') or not related (*pleurage* 'wow and flutter' with respect to *pleurer* 'to cry'), on a semantic and morphological basis; (ii) have no eventive meaning, such as *diction* 'diction' (from *dire* 'to say'). We also decided to limit the impact of polysemy. The models created by Word2Vec are not contextual and do not provide separate representations for the different meanings of polysemous words. In other words, forms like *garage*, meaning either the place where one parks a car or the action of parking, are represented by a single word embedding, aggregating the distributional information of the various meanings it covers. In order to limit the noise induced in the resulting representations, we excluded such nouns when we considered that the eventive meaning is less frequent than the non-eventive one – as is the case with *garage*. Finally, so as to ensure the coherence of the

data we use through this work, and because of the frequency threshold mentioned in Section 2.2., we systematically discarded the nouns with a frequency lower than 5 in the French *Wikipedia2018* corpus.

After the clean-up, we are left with 1.828 nouns, among which 629 are suffixed in *-age*, 750 in *-ion* and 449 in *-ment*. They are used to build the *-age*, *-ion* and *-ment* average vectors (henceforth AV_{AGE}, AV_{ION} and AV_{MENT}). Their first 50 nearest neighbors are given in Table 2. Neighbors which are not coined by the suffix of the class are highlighted in bold.

Tab. 2: 50 nearest neighbors of -age, -ion *and* -ment *average vectors* (AV_{AGE}, AV_{ION} *and* AV_{MENT}).

AV_{AGE}	*usinage* – 'machining' – *polissage* 'polishing' – *meulage* 'grinding' – *piquage* 'stitching' – *perçage* 'piercing' – *sablage* 'sand blasting' – *pliage* 'folding' – *remplissage* 'filling' – *salage* 'salting' – *soufflage* 'blow molding' – *démoulage* 'unmolding' – *décapage* 'scraping' – *sertissage* 'crimping' – *dégraissage* 'degreasing' – *assemblage* 'assembly' – *séchage* 'drying' – *rinçage* 'rinse' – **soudure** 'welding' – *enrobage* 'coating' – *nettoyage* 'cleaning' – **extrusion** 'extrusion' – *ponçage* 'sanding' – *compactage* 'compacting' – *broyage* 'crushing' – *malaxage* 'kneading' – *soudage* 'welding' – *façonnage* 'shaping' – ***recuit*** 'recooking' – *remontage* 'reassembly' – **rechargement** 'recharging' – *affûtage* 'sharpening' – *sciage* 'sawing' – *gonflage* 'inflation' – *tamisage* 'sieving' – *égouttage* 'draining' – *clouage* 'nailing' – **chargement** 'loading' – *calibrage* 'calibration' – *formage* 'shaping' – *brossage* 'brushing' – *réglage* 'tuning' – *traçage* 'tracing' – *cintrage* 'bending' – *lavage* 'washing' – *brasage* 'brazing' – *chromage* 'chrome plating' – *trempage* 'soaking' – *serrage* 'tightening' – *appareillage* 'casting off' – *coulage* 'pouring'
AV_{ION}	*généralisation* 'spread / generalization' – *manipulation* 'manipulation' – *dégradation* 'deterioration' – *simplification* 'simplification' – *stimulation* 'stimulation' – *contamination* 'contamination' – *dispersion* 'scattering' – *dénaturation* 'denaturation' – *transformation* 'transformation' – *récupération* 'recovery' – *utilisation* 'use' – *perception* 'perception' – *différenciation* 'differentiation' – *substitution* 'substitution' – *mutation* 'mutation / transfer' – *fixation* 'fastening' – *vérification* 'verification' – **persistance** 'persistence' – *prolifération* 'proliferation' – *assimilation* 'assimilation' – *altération* 'alteration' – *détermination* 'determination' – *réduction* 'reduction' – **saisie** 'seizure/input' – *dilution* 'dilution' – *conversion* 'conversion' – *activation* 'activation' – *compréhension* 'understanding' – *transmission* 'transmission' – *réutilisation* 'reuse' – *définition* 'definition' – **dégénérescence** 'degeneration' – **synthèse** 'overview / synthesis' – *redistribution* 'redistribution' – *modification* 'modification' – *multiplication* 'multiplication' – *régénération* 'regeneration' – *ponction* 'puncture' – **traitement** 'treatment' – *cristallisation* 'crystallization' – *décomposition* 'decomposition' – *détérioration* 'deterioration' – *stérilisation* 'sterilization' – *restriction* 'restriction' – *réaction* 'reaction' – *centralisation* 'centralization' – *dissociation* 'dissociation' – *dissémination* 'dissemination' – *standardisation* 'standardization' – *acceptation* 'acceptance'

AV_{MENT}	*déplacement* 'moving' – *étirement* 'stretching' – *durcissement* 'hardening' – *ajustement* 'adjustment' – *relâchement* 'slackening' – *traitement* 'treatment' – *adoucissement* 'softening' – **utilisation** 'use' – *échauffement* 'warm–up / heating' – *enfoncement* 'knocking in' – *décollement* 'detachment' – **rejet** 'rejection' – *rétrécissement* 'narrowing' – *affaiblissement* 'weakening' – *endommagement* 'damaging' – *fonctionnement* 'functioning' – *dépassement* 'passing' – *engorgement* 'congestion' – *allongement* 'lengthening' – *tassement* 'settling' – *encombrement* 'congestion' – *abaissement* 'lowering' – **usure** 'wear' – *débordement* 'overflowing' – *réajustement* 'readjustment' – *étalement* 'spread' – **remplissage** *'filling'* – *dégagement* 'release' – *écoulement* 'flow' – *éloignement* 'distancing' – **équilibrage** 'balancing' – *effritement* 'crumbling' – *isolement* 'isolation' – **inconfort** 'discomfort' – *fléchissement* 'bowing' – *gonflement* 'swelling' – *relèvement* 'raising' – *retournement* 'turnaround' – **déséquilibre** 'imbalance' – **colmatage** 'clogging' – *accroissement* 'increase' – *arrachement* 'abduction' – *basculement* 'toppling' – *accumulation* 'accumulation' – *balancement* 'swinging' – *amincissement* 'slimming' – **passage** 'passage' – *dessèchement* 'drying' – *dysfonctionnement* 'dysfunction' – **court–circuit** 'short circuit' – **activation** 'activation'

In the following, we conduct the analysis of their 100 nearest neighbors. This number is chosen arbitrarily so as to provide as broad a view as possible, and yet be significant.

2.4. Results analysis

These results show that the first 100 nearest neighbors of each average vector present a strong homogeneity: 82% of the AV_{AGE} neighbors are nouns in *-age*; 80% of the AV_{ION} neighbors are suffixed in *-ion*; 73% of the AV_{MENT} neighbors end with the *-ment* suffix. By comparison, the homogeneity is much lower for the neighbors of the average vectors of *-eur*, *-euse* and *-rice* agent nouns (Wauquier et al. 2018), with respectively 44%, 10% and 16% of their neighbors having the suffix of the class. In other words, AV_{AGE}, AV_{ION} and AV_{MENT} neighborhoods are located in distinct areas of the vector space, which reflects a clear distributional difference between the three classes of nominals.

On a semantic and referential level, the three neighborhoods exemplified in Table 2 almost exclusively contain action nouns but display some semantic differences. The AV_{ION} neighbors denote processes or phenomena related to sciences, such as *dilution* 'dilution' or *dénaturation* 'denaturation', which goes in line with Dubois' claim. There are also nouns describing psychological processes, such as *compréhension* 'understanding', *détermination* 'determination', or *perception* 'perception', and to a greater extent nouns denoting very broad concepts, such as *modification* 'modification' and *utilisation* 'use', characterized by a high degree of polysemy. The neighborhood of

the AV_{AGE} vector contrasts strongly with that of AV_{ION}. It contains a large amount of nouns related to industrial skills or processes, such as *soudage* 'welding', *usinage* 'machining', and *brasage* 'brazing', and very few, if not any, generic or underspecified nouns. The most generic nouns like *stockage* 'storage', *nettoyage* 'cleaning' and *lavage* 'washing' denote actions which are intrinsically more technical than *utilisation* or *modification* and many other neighbors of the AV_{ION} vector. The AV_{MENT} neighborhood seems more mixed. Some neighbors designate relatively large concepts, such as *déplacement* 'moving' and *traitement* 'treatment', while others are more specific, such as *relèvement* 'raising' and *équilibrage* 'balancing'.

To summarize, the two semantic profiles that emerge for the nominals in *-age* and *-ion* are clearly distinct: the former displays a higher degree of technicality, while the latter has a higher degree of genericity. The semantic profile of the nominals in *-ment* is less marked. In what follows, we focus on this distinction and we propose a set of criteria that characterize technicality.

3. The concept of technicality

Technicality is a notion hardly ever discussed in the literature. The adjective 'technical' is sometimes used to refer to specific types of vocabulary or corpora, but technicality has not yet been properly defined, as Mudraya (2006) points out. However, we need such a definition in order to better characterize the semantic distinction we just highlighted. To the best of our knowledge, no work has been dedicated to this notion of technicality from a non-terminological point of view. Conversely, outside the domain of linguistics, the philosopher Simondon (1958) extensively examines the issue. First we review his description, then we propose our own definition of technicality for action nouns. Finally, we investigate its implementation by means of linguistic criteria that can be automatically measured. It should be pointed out that we do not take a terminological stand on technicality here. Our aim is to propose a definition that relies on lexical and referential criteria and is valid independently of any particular domain.

3.1. Definition of technicality

Simondon (1958) gives a definition of technicality along several dimensions. The first one is agentivity. According to the author, technicality is closely linked to human beings, as they are "among the machines that operate with them" ("*[ils sont] parmi les machines qui opèrent avec [eux]*") (1958: 12). Incidentally, the philosopher defines technicality with respect to machines and more largely to objects he describes as technical: "technicality manifesting itself through the use of technical objects" ("*la technicité*

se manifestant par l'emploi d'objets techniques") (1958: 156). The degree of technicality of an object depends on its level of sophistication and complexity, estimated according to whether the knowledge required to use it is innate or acquired. An innate, unreflective knowledge, related to an everyday life object, conveys a lesser degree of technicality of the action, whereas an action that involves "a reflective operation, a rational knowledge constructed by sciences" ("*une opération réfléchie, d'une connaissance rationnelle élaborée par les sciences*") (Simondon 1958: 85) is considered technical. The more the knowledge is learned and constructed, the higher the technicality of the denoted action. Ontologically, Simondon indicates that "technique concerns business, agriculture, industry" ("*la technique touche au commerce, à l'agriculture, à l'industrie*") (Simondon 1958: 97). Following Simondon (1958), we propose in (1) the definition of technical action nouns.

(1) A technical action noun is a noun unfamiliar to non-experts, denoting a specific and complex action, whose achievement and understanding require an acquired skill and which is specific to a particular domain. Technical action nouns typically, but not exclusively, belong to domains such as industry, agriculture and arts and crafts.

As such, we consider as technical any noun that denotes an action performed intentionally by an agent and whose complexity requires from the agent some specific learned knowledge that might (but not necessarily) be at the core of a well-defined task. Technical actions cannot be successfully carried out by all comers in everyday life, as they involve in addition to specific knowledge a particular setup. This distinguishes action nouns such as *danse* 'dance' and *ébreuillage* 'fish evisceration': one can dance at any time even while not being a professional dancer, while fish evisceration requires particular skills and gears to be performed correctly. The complexity and specificity of the action result in the unfamiliarity of non-experts with the corresponding nouns. According to our approach, domains other than industry or agriculture, such as scientific disciplines, can also provide technical nouns that denote complex actions involving tools and devices, as well as specific knowledge (e.g. *inoculation* 'inoculation', *cassation* 'quashing'). However, being part of a specialized domain is neither a necessary nor a sufficient condition for a noun to be technical (see *aimance* 'child-specific affection', which belongs to a specialized domain, i.e. the psychology field, yet does not meet our definition of technical noun). Note that while definition in (1) suggests that there is a class of technical nouns, technicality is actually considered as a gradable property that is instantiated at various degrees.

3.2. Linguistic properties of technicality

As we just saw, technicality relies on world knowledge and referential criteria and needs to be translated into linguistic properties that can be empirically investigated. Three linguistics properties of technicality (henceforth *T1*, *T2* and *T3*) can be derived from the previous definition and implemented by a set of criteria automatically calculated on corpora and lexicons.

The first is **specialization** (*T1*): a technical action noun is more often used in specialized contexts than in general contexts because it falls within a particular domain and the action it denotes is specific. This denotational specificity gives rise to the second property, namely **obscurity** (*T2*): a technical action noun is more often described and explained to non-experts than a non-technical noun because it denotes a reality which is unfamiliar to them. The description can be a dictionary definition, an article in an encyclopedia, etc. The third property we derive from the definition in (1) is **univocity** (*T3*): a technical action noun tends to be unequivocal, in contrast with a generic noun. Monosemy can be used as an approximation of univocity.

3.3. Criteria to approximate technicality

At this stage, the previous three linguistic properties must be operationalized. Several measures can be used to assess the degree of specialization, obscurity and univocity of a given action noun. Our choice is influenced by the need for automatic annotation and is based on available linguistic resources. These criteria are exploratory and we are aware that this first attempt to estimate technicality will have to be refined with respect to the preliminary results we present in Section 4. Table 3 presents the criteria that we derived from the properties just described.

Tab. 3: Technicality criteria.

Property	Criteria
T1	Ratio of the relative frequencies in a technical corpus and in a reference corpus
	Number of lexicographical markers of domains in dictionaries or encyclopedia
	Presence or absence in transdisciplinary scientific lexicons
T2	Presence or absence of an article in an encyclopedia
T3	Number of synonyms
	Number of definitions in dictionaries
	Presence or absence in generic nouns lexicons

The degree of specialization of a word is assessed by corpora comparison (Lemay et al. 2005) and by means of lexicons that provide information on

domain membership and transdisciplinarity (Hatier 2016). As for the obscurity property, we use one criterion: the presence or absence of the noun as an entry in an encyclopedia. Different *T3* criteria allow for the approximation of the various aspects of equivocality, namely polysemy and underspecification. These criteria are computed from various resources, as shown in Table 4.

Tab. 4: Resources and lexicons.

Name	Size	Description
Wikipedia2018	600 million words	Encyclopedic corpus built from the 2018 French edition of *Wikipedia*
LM10	200 million words	French journalistic corpus made up of articles of *Le Monde* newspaper published between 1991 and 2000
DES	83,395 entries	Electronic dictionary of synonymes (Manguin et al 2004)
TLFi	54,280 entries	Electronic version of the *Trésor de la Langue Française* dictionary (Dendien and Pierrel 2003)
GLAWI	1,481,346 entries	Electronic dictionary built from the French *Wiktionary* (Hathout and Sajous 2016)
LexiTrans	1,611 entries	Transdisciplinary scientific lexicon (Drouin 2010)
LexNSS	305 entries	List of underspecified nouns extracted from Legallois and Gréa (2006)

In this study we choose to use several dictionaries in order to reduce the impact of specific lexicographic choices. Regarding corpora, the *LM10* corpus has been chosen as a reference corpus, as it is considered to be an example among others of non-technical discourse. The choice of the *Wikipedia2018* corpus as a technical corpus is supported by its encyclopedic nature, by the fact that it incorporates a large panel of technical domains, but also by the lack of a large and diversified technical corpus for French. We also use *Wikipedia2018* to test the presence or absence of an article describing the action denoted by the noun (PAGE_W18). Only the articles with a title strictly identical to the noun are taken into account. For example, we consider that *serrage* 'tightening' does not have an article in *Wikipedia2018* even if it contains articles entitled *collier de serrage* 'horse clamp' or *noix de serrage* 'clamp holder'. As for lexicons, we both use large general dictionaries for French (*Trésor de la Langue Française* and *Glawi*) and smaller specific lexicons that give access to synonymy, transdisciplinary vocabulary (as a clue for non-specialization) and shell nouns (as a clue for equivocality).

Our criteria along with their associated resources are given in Table 5.

Tab. 5: Implementation of the technicality criteria.

Property	Criteria	Name
T1	Ratio of the relative frequencies (per million words) in *Wikipedia2018* and *LM10*	RATIO_FREQR
	Number of category markers in *Wikipedia2018*	NB_CAT_W18
	Number of lexicographical markers of domains in *TLFi*	NB_DOM_T
	Number of lexicographical markers of domains in *GLAWI*	NB_DOM_G
	Presence or absence in *LexiTrans*	LST
T2	Presence or absence of an article in *Wikipedia2018*	PAGE_W18
T3	Number of synonyms in *DES*	NB_SYN
	Number of definitions in *TLFi*	NB_DEF_T
	Number of definitions in *GLAWI*	NB_DEF_G
	Presence or absence in *LexNSS*	NSS

As we can see in this table, we use very simple measures that check for the presence of the noun in the lexicons or count the number of lexical items (definitions, synonyms) that are found in the entry. The ratio of relative frequency is calculated by dividing for a given word its relative frequency in the technical corpus by its relative frequency in the reference corpus (Hatier 2016).

It should be noted that technicality is estimated by the combination of these criteria. They aim to highlight tendencies in the degrees of technicality of action nouns, and as such, there is no threshold value we could use as a clue for a binary characterization of a noun as technical or non-technical.

4. Statistical modeling of technicality

The criteria we just presented enable us to empirically test the hypothesis of a higher degree of technicality for *-age* nouns and of a lesser degree of technicality for *-ion* and *-ment* nouns. Following the definition of technicality given in Section 3.1., *-age* nominalizations are expected to have higher values than *-ion* and *-ment* nominalizations for only two criteria related to *T1* and *T2*, and lower values than *-ion* and *-ment* nominalizations for the others criteria: they will have a higher frequency ratio (RATIO_FREQR) and will be more likely to have an article in Wikipedia2018 (PAGE_W18), but they will be less present in the transdisciplinary lexicon (LST) and among the underspecified nouns (NSS), they will have fewer synonyms (NB_SYN), definitions (NB_DEF), and domain markers (NB_DOM).

4.1. Annotation

To test our predictions, we automatically annotated the 1828 *-age*, *-ion* and *-ment* action nouns selected in Section 2.1. for the criteria presented in Section 3.3. Table 6 presents the annotation of 4 nouns selected to illustrate the opposition between technical nouns (*alunissage* 'moon landing' and *cimentage* 'cementing') and non-technical nouns (*correction* 'correction' and *revendication* 'demand') according to our definition in (1). Our predictions are shown in the table with the (+) and (–) marks. First, we can see that the 4 examples are fairly well described by the criteria we implemented even if some criteria, when considered individually, do not systematically conform to our expectations. For example, the absence of a noun in the transdisciplinary (LST) lexicon does not ensure its technicality (*revendication*). Moreover, a high number of definitions (NB_DEF_G and NB_DEF_T) is not necessarily a good clue for the non-technicality of the noun (*cimentage*). Yet, the technical nouns *alunissage* and *cimentage*, denoting respectively a specific maneuver of a spatial engine and a process of the construction industry, have several values that are close to our expectations (higher ratio in RATIO_FREQR, and fewer synonyms, definitions and domains, respectively in NB_SYN, NB_DEF and NB_DOM). Similarly, non-technicality seems to be globally captured, as shown by the nouns *correction* and *revendication*.

Tab. 6: Values of the technicality criteria for the action noms alunissage *'moon landing'*, cimentage *'cementing'*, correction *'correction' et* revendication *'demand'*.

	Technicality	***alunissage***	***cimentage***	***correction***	***revendication***
RATIO_FREQR	+	3.06	2.18	1.14	0.23
NB_CAT_W18	–	3	0	0	0
NB_DOM_T	–	1	2	7	3
NB_DOM_G	–	1	0	4	1
LST	–	No	No	No	No
PAGE_W18	+	Yes	No	Yes	No
NB_SYN	–	1	0	87	45
NB_DEF_T	–	1	3	33	6
NB_DEF_G	–	1	1	8	3
NSS	–	No	No	No	Yes

Table 6 highlights the role of our criteria as tendency indicators and not as class delimiters. We do not have technical and non-technical nouns *per say*, but nouns that have a higher degree of technicality than others. Among the nouns having the highest degree of technicality with respect to our criteria, i.e. the nouns that overall conform the most to our predictions, we find *hydroformage* 'hydroforming', *zingage* 'galvanizing', *cardage* 'carding' and *oxycoupage* 'oxy cutting', and among the lowest *association* 'combination', *division* 'division', *commencement* 'beginning' and *approbation* 'endorsement'.

We provide in Table 7 the average values of the technicality criteria for the 1.828 action nouns in *-age*, *-ion* and *-ment*. The presence in LST and NSS lexicons is presented as the percentage of nouns belonging to these lexicons. The existence of an article corresponding to the noun in the *Wikipedia2018* corpus is also given as a percentage.

Tab. 7: Average values of the technicality criteria with respect to each suffix.

	-age	*-ion*	*-ment*
RATIO_FREQR	2.62	2.16	1.57
NB_CAT_W18	0.78	0.93	0.42
NB_DOM_T	1.2	2.65	1.27
NB_DOM_G	0.65	0.93	0.46
LST (%)	0.1	8.13	2.23
PAGE_W18 (%)	48.97	69.87	34.08
NB_SYN	3.03	13.83	11.01
NB_DEF_T	2.52	5.8	4.34
NB_DEF_G	2.01	2.54	1.98
NSS (%)	0	2	1.11

The results shown in Table 7 corroborate the hypothesis of a higher degree of technicality of *-age* action nouns, and a lesser degree of technicality of *-ion* action nouns. We can see that *-age* nominalizations have on average fewer synonyms (NB_SYN), definitions (NB_DEF) and domain markers (NB_DOM) than *-ion* nominalizations (p-value < 0.01 except for the number of category markers in *Wikipedia2018*, i.e. NB_CAT_W18). Moreover, they are proportionally less present in the LST and NSS lexicons than *-ion* nominalizations (p-value < 0.01). It can be noticed that the differences of values displayed by the suffixes are higher for measures extracted from *TLFi* (NB_DEF_T and NB_DOM_T) than for those from *GLAWI* (NB_DEF_G and NB_DOM_G).

However, one criterion is not in line with our predictions, namely PAGE_W18 (presence or absence of a page in *Wikipedia2018*). We can see in Table 7 that the percentage of nouns having an article in the Wikipedia2018 corpus is lower for *-age* than for *-ion* (49% for the former vs 70% for the latter). Although the difference between each pair of suffixes is significant (p-value < 0.01), it fails to approximate the technicality of action nouns. Moreover, while the average ratio of relative frequency (RATIO_FREQR) is slightly higher for *-age* than for *-ion*, the difference is not significant. The inability of these criteria to discriminate between technical and non-technical action nouns is probably explained by the choice of the *Wikipedia2018* corpus. It is not technical enough and consequently not appropriate to evaluate the properties regarding specialization and obscurity.

Table 7 also shows that *-ment* action nouns are in an intermediate position with respect to our criteria: we notice that its average values lie between those of *-age* and *-ion* for criteria such as the number of synonyms (NB_SYN), the number of definitions in *GLAWI* (NB_DEF_G) or the presence in LST and NSS lexicons. Yet, its average values for other criteria such as the number of categories in *Wikipedia2018* (NB_CAT_W18) or the number of domains markers in both dictionaries (NB_DOM) are lower than those of *-age*. Furthermore, the difference between *-age* and *-ment* is not significant for NB_DEF_G and NB_DOM_T. In contrast, the average values for the suffix *-ment* are never higher than those of *-ion*. These observations suggest that *-ment* nominalizations are closer to *-age* than to *-ion* with respect to the technicality of the actions they denote.

4.2. Predictive capability of the criteria

In the last step of our analysis, we can now assess the extent to which our technicality criteria allow for the prediction of the suffix of an action noun. Their discriminative capability can be estimated by means of a decision tree which predicts the suffix of the noun from the values of the criteria. We used the *rpart* package in R to build a classification decision tree that classifies the 1828 *-age*, *-ion* and *-ment* action nouns based on their annotation. The resulting decision tree is presented in Figure 1. The values given under the leaves are the number of *-age*, *-ion* and *-ment* nominalizations respectively in each cluster. The letters "O" and "N" respectively stand for *yes* and *no*.

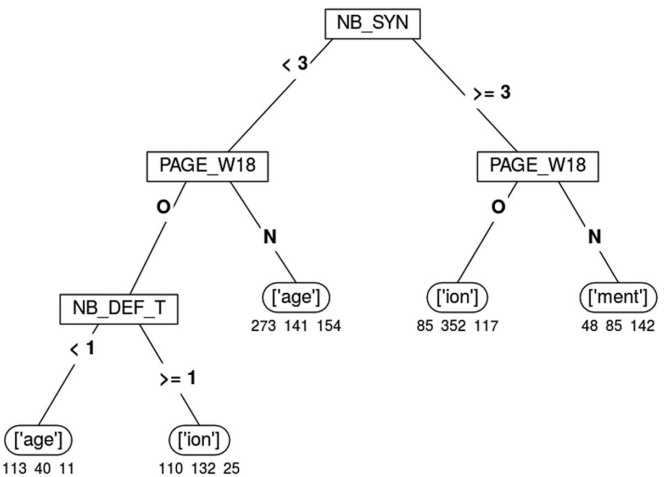

Fig. 1: Classification tree.

Figure 1 shows both the rules inferred from the data and the contribution of the criteria to the classification. We see that only 3 of the 10 initial criteria

are actually used to classify the nominals, namely the number of synonyms (NB_SYN), the number of definitions in *TLFi* (NB_DEF_T) and the presence or absence of an article in Wikipedia2018 (PAGE_W18). This last rule confirms the observation made from Table 7 that this criterion does not satisfactorily operationalize the *T2* property and shows that it actually plays in the opposite way. The PAGE_W18 criterion selects the non-technical nominals, and the *-ion* suffix.

The overall accuracy of the model is 55.3%, meaning that hardly half of the action nouns are properly classified. Yet the three suffixes are not classified with the same precision. The performances of the model are given in the confusion matrix in Table 8. The number of items predicted as *-age*, *-ion* and *-ment* nominals (in the columns) is given with respect to the actual number of these nominals (in the rows). The correctly classified items are highlighted (in bold). Table 8 shows that 386 nominals in *-age* are identified as part of the *-age* class, 195 as *-ion* nouns and 48 as *-ment* nouns. It also shows that the model achieves similar performances for *-age* and *-ion* (respectively 61.4% and 64.5%) but a lower performance for *-ment* (31.6%).

Tab. 8: Confusion matrix.

		Predicted		
		-age	*-ion*	*-ment*
Observed	*-age*	**386**	195	48
	-ion	181	**484**	85
	-ment	165	142	**142**

A closer look at the misclassified nouns first shows that the criteria NB_SYN, NB_DEF_T and PAGE_W18 do a good job in discriminating nouns in terms of technicality. Among the misclassified *-age* nominals (i.e. the ones that are predicted to be *-ion* or *-ment* nouns), we find nouns such as *papotage* 'chattering', *batifolage* 'frolic' and *babillage* 'babbling' which are not technical and are labeled as *-ment* by the classifier due to their high number of synonyms. Although they are misclassified at the morphological level, the prediction is correct on the semantic level, given their low degree of technicality. Other misclassified nouns such as *damage* 'tamping', labeled as *-ion* because it has a definition in *TLFi*, show that some rules are given too much importance. However, the error analysis brings to light that 11% of the misclassified *-ion* nouns (i.e. nominals in *-ion* which were assigned to another suffix) are suffixed in *-ification* or *-isation* (*panification* 'breadmaking'; *dessalinisation* 'desalination') and labeled as *-age* because of the inherent technicality of their base verbs.

In a second experiment, we further refined our assessment of the criteria. We excluded the PAGE_W18 criterion because it does not properly implement the *T2* property. The resulting decision tree is given in Figure 2.

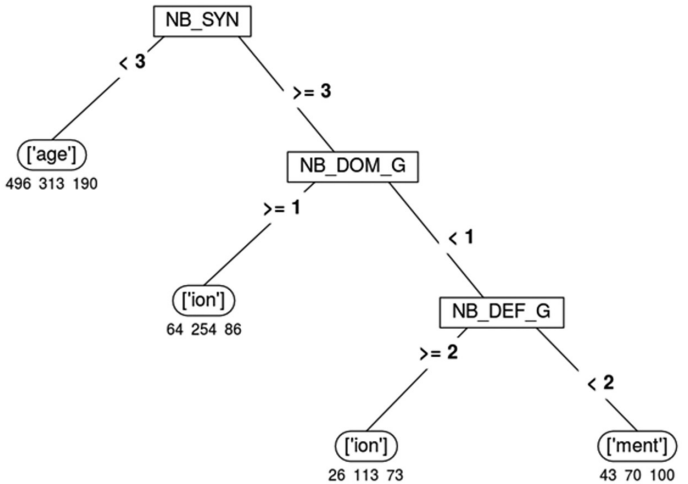

Fig. 2: Decision tree excluding the PAGE_W18 criterion.

In the new model the number of synonyms is still the first criterion and is now followed by the number of domains and the number of definitions in *GLAWI*. The rules inferred by this model are more in line with our hypothesis and predictions: *-age* nominals are discriminated from the others by their lower number of synonyms (< 3); action nouns are labeled as *-ion* nominals if they have more than one domain marker in *GLAWI*; *-ion* nominals are discriminated from the *-ment* nominals by their number of definitions (>=2). Compared to the previous model, this one has a lower overall accuracy (52.7%), but has an improved precision for *-age* nominals (78.9%) and a lower one for *-ion* (48.9%) and *-ment* (22.3%). The corresponding confusion matrix is given in Table 9.

Tab. 9: Confusion matrix.

		Predicted		
		-age	*-ion*	*-ment*
Observed	*-age*	**496**	90	43
	-ion	313	**367**	70
	-ment	190	159	**100**

A closer look at the *-age* misclassified nouns shows that they were already misclassified by the previous model. The labeling of *-ification* and *-isation*

nominalizations as technical is emphasized, since 53.2%, respectively 81%, of them are considered to be *-age*, vs 10.6%, respectively 7.1%, as *-ment*.

This experiment challenges the relevance of some of our criteria, in particular those linked to the corpora. Even though it was used in an opposite direction, PAGE_W18 actually improves the overall prediction of the first model, and in particular the classification of the *-ion* and *-ment* nouns. Regarding the ratio of relative frequencies, it may be that the corpus *Wikipedia2018* is not technical enough, and too diverse to actually favor *-age* nominalizations. Regarding the presence of an article in *Wikipedia2018*, we might want to revise its implementation and the hypothesis supporting this criterion. Taking into account all the articles whose title contains the noun might emphasize more the genericity than the technicality. While articles such as *noix de serrage* 'clamp holder' would be taken into account, allowing for *serrage* 'tightening' to validate the PAGE_W18 criterion, many generic nouns would also fulfill the PAGE_W18 criterion, such as *utilisation* 'use' which occurs in *utilisation frauduleuse des instruments de paiement* 'fraudulent use of payment.

5. Conclusion

In this study, we investigated the factor of technicality in rivalry between French action nouns in *-age*, *-ion* and *-ment*. We first proposed a definition of technical action nouns as being nouns unfamiliar to non-experts, which denote a specific, complex action that requires an acquired skill and is grounded in domains such as industry, agriculture and handicraft. We showed how linguistic properties and criteria could be extracted from this definition and computed from corpora and lexical resources in order to characterize the technicality of the nominals. Some of these criteria, such as the number of synonyms and definitions, have proven to be effective in discriminating *-age* nominals which are more technical, from *-ion* nominals which are less technical. Other criteria turned out to be less relevant, in particular the ones computed from corpora (e.g. the ratio of relative frequencies).

We also saw that *-ion* nominals are more heterogeneous than expected in terms of technicality, mainly because of the presence of inherently technical *-isation* and *-ification* action nouns. In future work, we intend to use a finer-grained dataset where these nominals will be part of the same class as *-age* nominals. We will also take into account other derivational processes, such as conversion (*baisse* 'lowering') or *-ure* suffixation (*rature* 'crossing out').

This work is a first attempt to characterize and approximate technicality through empirical criteria. It offers a new insight on the distinction between *-age*, *-ion* and *-ment* French nominals and highlights the limits of some exploratory criteria. In particular, we need to improve the criteria that estimate obscurity. We also intend to explore new criteria such as concreteness (Pierrejean 2020; Köper & Schulte im Walde 2016) and instrumentality (Missud 2019):

there is a certain amount of overlap between technicality and the simpler notion of concreteness because technical action nouns mostly denote concrete actions (*usinage* 'machining', *extrusion* 'extrusion'), as opposed to abstract ones (*compréhension* 'understanding', *perception* 'perception'). Finally, we will also evaluate the possible correlation between our criteria and the level of technicality perceived by the speakers. Ultimately, we aim to propose a single technicality score that aggregates the relevant criteria and annotations.

References

Dendien, Jacques & Jean-Marie Pierrel. 2003. Le Trésor de la Langue Française Informatisé: un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des Langues* 44(2). 11–37.

Drouin, Patrick. 2010. Extracting a Bilingual Transdisciplinary Scientific Lexicon. In Sylviane Granger & Magali Paquot (eds.), *eLexicography in the 21st century: new challenges, new applications. Proceedings of eLex 2009*, 45–53. Louvain-la-Neuve: Presses Universitaires de Louvain.

Dubois, Jean. 1962. *Étude sur la dérivation suffixale en français moderne et contemporain: essais d'interprétation des mouvements observés dans le domaine de la morphologie des mots construits*. Paris: Larousse.

Firth, John R. 1957. A synopsis of linguistic theory 1930–1955. In John R. Firth (eds), *Studies in linguistic analysis*, 1–32. Oxford: Basil Blackwell.

Fleischman, Suzanne. 1990. *The French Suffix -age: its Genesis, Internal Growth, and Diffusion*. Ann Arbor: Univ. Microfilm Intern.

Fradin, Bernard. 2014. La variante et le double. In Florence Villoing, Sarah Leroy & Sophie David (eds.), *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*, 109–147. Paris: Presses universitaires de Paris Nanterre.

Harris, Zellig S. 1954. Distributional structure. *Word* 10(2-3). 146–162.

Hathout, Nabil & Franck Sajous. 2016. Wiktionnaire's Wikicode GLAWIfied: A Workable French Machine-Readable Dictionary. In *Proceedings of the tenth International Conference on Language Resources and Evaluation (LREC)*, 1369–1376. Portorož.

Hatier, Sylvain. 2016. *Identification et analyse linguistique du lexique scientifique transdisciplinaire. Approche outillée sur un corpus d'articles de recherche en SHS*. Grenoble: University Grenoble Alpes doctoral dissertation.

Kelling, Carmen. 2001. Agentivity and Suffix Selection. In *Proceedings of the Lexical-Functional Grammar '01 Conference (LFG'01)*, 147–162. Hong Kong.

Kintsch, Walter. 2001. Predication. *Cognitive science* 25. 173–202.

Köper, Maximilian & Sabine Schulte im Walde. 2016. Automatically Generated Norms of Abstractness, Arousal, Imageability and Valence for 350 000 German Lemmas. In *Proceedings of the 10th Conference on Language Resources and Evaluation (LREC), 2595–2598*. Portorož.

Legallois, Dominique & Philippe Grea. 2006. L'objectif de cet article est de…Construction spécificationnelle et grammaire phraséologique. *Cahiers de praxématique* 46. 161–186.

Lemay, Chantal, L'Homme, Marie-Claude & Patrick Drouin. 2005. Two methods for extracting "specific" single-word terms from specialized corpora: Experimentation and evaluation. *International Journal of Corpus Linguistics* 10(2). 227–255.

Lenci, Alessandro. 2018. Distributional Models of Word Meaning. *Annual review of Linguistics* 4. 151–171.

Manguin, Jean-Luc, Jacques François, Rembert Eufe, Ludwig Fesenmeier, Corinne Ozouf & Morgane Senechal. 2004. Le dictionnaire électronique des synonymes du CRISCO: un mode d'emploi à trois niveaux. *Cahiers du CRISCO* 17.

Marelli, Marco & Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological review* 122(3). 485–515.

Martin, Fabienne. 2010. The Semantics of Eventive Suffixes in French. In Artemis Alexiadou & Monika Rathert (eds.), *The Semantics of Nominalizations across Languages and Frameworks*, 109–141. Berlin & New York: de Gruyter.

Mikolov, Tomas, Kai Chen, Greg Corrado, & Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of International Conference on Learning Representations (ICLR)*. Scottsdale.

Missud, Alice. 2019. *Modélisation quantitative de la rivalité entre la suffixation en -age et la conversion de verbe à nom*. Paris: University Paris Nanterre master thesis.

Mudraya, Olga. 2006. Engineering English: A lexical frequency instructional model. *English for Specific Purposes* 25(2). 235–256.

Pierrejean, Bénédicte. 2020. *Qualitative Evaluation of Word Embeddings: Investigating the Instability in Neural-Based Models*. Toulouse: University Toulouse - Jean Jaurès doctoral dissertation.

Simondon, Gilbert. 1958. *Du mode d'existence des objets techniques*. Paris: Éditions Aubier-Montaigne.

Urieli, Assaf. 2013. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Toulouse: University Toulouse - Jean Jaurès doctoral dissertation.

Uth, Melanie. 2010. The rivalry of French -ment and -age from a diachronic perspective. In Artemis Alexiadou & Monika Rathert (eds.), *The Semantics of Nominalizations across Languages and Frameworks*, 215–244. Berlin & New York: de Gruyter.

Wauquier, Marine, Cécile Fabre & Nabil Hathout. 2018. Différenciation sémantique de dérivés morphologiques à l'aide de critères distributionnels. In *Congrès Mondial de Linguistique Française (CMLF)*, Mons, Belgique.

Marine Wauquier, Nabil Hathout & Cécile Fabre
CLLE, University Toulouse Jean Jaurès & CNRS
5, allée Antonio Machado
31058 Toulouse cedex 9
France
{marine.wauquier, nabil.hathout, cécile.fabre}@univ-tlse2.fr