Methodology to build and exploit a representative corpus for neological study in the field of medicine

Coralie Schneider
Universitat Pompeu Fabra
Coralie.schneider@upf.edu

Rosa Estopà
Universitat Pompeu Fabra
rosa.estopa@upf.edu

Coralie Schneider és investigadora en Lingüistica de Corpus i Terminologia del grup IULATERM de la Universitat Pompeu Fabra. Ha defensat la seva tesi doctoral en Lingüística en cotutela a la Universitat Pompeu Fabra i a la Universitat París Diderot, i és alumna de l'École Normale Supérieure de Cachan. Ha impartit docència de francès a la Universitat d'Oxford. Centra la seva recerca en terminologia i fraseologia en el discurs mèdic.

Rosa Estopà és professora titular de la Universitat Pompeu Fabra, on ha impartit docència en terminologia, neologia, lexicografia, lingüística aplicada, comunicació mèdica i català des de 1994. Llicenciada en Filologia Catalana (Universitat de Barcelona, 1992), Màster en Patologia del llenguatge i ciències de l'audició (Universitat Autònoma de Barcelona, 1993) i doctora en Lingüística (Universitat Pompeu Fabra, 1999). És investigadora de l'Institut de Lingüística Aplicada (IULA), del grup IULATERM i de l'Observatori de Neologia. Coordina el Màster Online en Terminologia de la UPF.

Resum

Aquest article pretén introduir una nova metodologia dissenyada per construir un corpus adequat per a l'estudi de fenòmens neològics (patrons de creació neològica, formació de termes primaris o secundaris, categories de conceptes a què fan referència, difusió neològica, existència de variants potencials...). Això ens va permetre analitzar les probabilitats d'integració neològica del llenguatge mèdic a partir dels factors de supervivència neològica *que anteriorment vam identificar i estudiar.*

Paraules clau: *lingüística de corpus; detecció de neologia; metodologia; neologismes especialitzats; diacronia; terminologia mèdica*

Abstract

This paper introduces a new methodology designed to build an adequate corpus for the study of neological phenomena (neological creation patterns, primary or secondary term formation, concept categories they refer to, neological dissemination, existence of potential variation...). This allowed us to analyse the probabilities for neological integration in the medical language in the light of the neological survival factors *we previously identified and studied.*

Keywords: *corpus linguistics; neology detection; methodology; specialised neologisms; diachrony; medical terminology*

Terminàlia 22 (2020): 29-39 · DOI: 10.2436/20.2503.01.154
Data de recepció: 24/09/2020. Data d'acceptació: 28/10/2020
ISSN: 2013-6692 (impresa); 2013-6706 (electrònica) · http://terminalia.iec.cat

1 Introduction and objectives

We introduce a new methodology designed to build an adequate corpus for neology study. This allows to analyse the probabilities for neological integration in a specialised language in the light of the *neological survival factors* which we previously identified and studied (Schneider, 2018).

This paper has been divided in seven sections. First, the main goals are presented. The type and source of the literature to be compiled constitute the second section. The third section is dedicated to identifying, extracting specialised neologisms upon which the research relies. The fourth section stems from the necessity to integrate in the corpus, a representative (ideally an exhaustive) part of the literature published on each of the concepts referred to by the specialised neologisms. The fifth section is devoted to the resulting exploitable corpus for the study of neologisms. The benefits and the limitations of this novel methodology are discussed in section six. Finally, the results that this methodology has enabled to obtain are discussed and new avenues for further data collection are suggested.

1.1 Specialised neology

Specialised neology may be defined as a *terminological status* or a phase in a term's life, spanning from its "birth" (that is from the first time it appeared) to the time it either became an established term or a discarded neologism. Defining exactly and pragmatically when a neologism ceases to be is unrealistic. Linguists such Auger (2010), Cabré (1998, 2006), Humbley (2003), Quirion (2010), Rey (1976), Sablayrolles (1996, 2000) and Sager (1989) have provided definitions which lead to consider that a neologism remains as such while it is still perceived as "new" within a linguistic community and that this *feeling of novelty* could remain for several years. Thus it is paramount that the research corpus cover several years (ideally a longer period than what is needed for the *feeling of novelty* to disappear). Whereas it might be difficult to detect terms that fully abide by the conditions to be neologisms in the present time, it is judicious to study them in context from the time of their first occurrence in the past until today, to be sure to integrate the lapses of time during which they were neologisms.

1.2 Neological survival factors

In order to further the study of specialised neologisms, we identified a series of quantitative and qualitative criteria (or factors) which can be held accountable for the integration of newly created terms or for their disappearance. We referred to them as *neological survival factors*:

- The use frequency for each neologism within a linguistic community.
- The distribution rate (that is the number of different publications in which a neologism occurs at least once).
- The presence of terminological variants. This synonymy situation may happen when several research teams are studying the same new concepts and decide, each on their own, to create new terms to refer to them. It may also happen when a previous term is no longer considered appropriate to refer to a concept. The presence of a terminological variant may have an influence on the evolution and integration probabilities of the other variants.

As for terminological variants, it is necessary to study the diachronic evolution of use frequencies and distribution rates in order to monitor their relative importance. For each occurrence of each neologism or their potential variants, an in-depth study of the context is necessary:

- The co-text where neologisms and terminological variants occur (that is, the collocations of words directly on the right or on the left of a given term).
- The specific locations where neologisms and terminological variants occur within a publication. A neologism appearing in a title, an introduction, a keyword list or a conclusion, should theoretically benefit from a greater visibility to the reader, than if it only appeared in the main text.
- The type of publications and especially their level of specialisation should be taken into account when analysing the context in which a neologism occurs.
- The influence of key reference articles regarding the concept that is referred to by a neologism or a variant. We determined their degree of influence by taking into account the number of times that an article, whose main title contains either a neologism or a variant is quoted in later publications.

1.3 Neological corpus' characteristics

To analyse the evolution of each survival factor, it is paramount to define the characteristics of the research corpus. These characteristics will allow to select the type of texts to be included in the corpus. The following considerations (marked in bold) formulated by Cabré (1998) allow to compile a research corpus. **The goal to be achieved** is to detect medical neologisms (i.e. medical terms which appeared for the first time during the period covered by the corpus) and for each of them, gather data on each factor that may influence their integration into the medical language. **The types of data that need to be collected in the corpus in order to reach this goal:** quantitative and qualitative data for each studied neologism, as well as for each of their terminological variants. Variants may end up becoming established terms to the detriment of the studied neologism. Quantitative data refers here to use frequen-

cy and distribution rate. Qualitative data means here the co-text, the context, the specific locations where a neologism occurs, and the influence of key reference articles or authors. **The criteria that the corpus must meet to enable to achieve this objective**: in order to monitor the yearly evolution of each survival factor, it is necessary to build a diachronic corpus, subdivided per year, which covers several years (ideally more than a decade). The longer the period covered, the easier it will be to determine the trend followed by each survival factor's evolution and therefore the easier it will be to predict the neologism's integration or demise based on whether the trend follows a growing or a declining curve (Rey, 1972). The next task consists in selecting the specialised literature to integrate in the research corpus.

2 Preliminary corpus compilation

2.1 Determining which specialised literature to integrate

To build the research corpus, it is necessary to define the studied field. The definition should contain discriminatory criteria to select articles published in that field only. For instance, the medical subfield of rare diseases has been defined differently and multiple times by medical researchers and institutions. The definition adopted by the European Union – a rare disease affects less than one person out of 2000 – is quantifiable and excludes all diseases with a higher prevalence. Therefore a list of diseases whose prevalence is inferior to 0.005, created from the Orphanet database (INSERM, 1997) was drawn in order to select articles dealing with them.

Moreover, considering that rare diseases prevalence may differ from one part of the world to another, it is also necessary to take into account any disease considered as "*rare*" or "*very rare*" or "*ultra rare*" by the author. It is indeed probable that they may include neologisms related to this field.

2.2 Where to look for the texts to integrate

Queries are to be carried on *Google*, *Google Scholar*, *ScienceDirect*, *Pubmed* (or any other relevant scientific platforms, databases or corpora such as the *British National Corpus*, *Frantext* or *Leipzig Corpora Collection*) to collect research articles meeting field-specific, discriminatory criteria. For instance, in our research, the articles have to deal with diseases from the rare-diseases list. Nine English-language journals specialised in rare diseases were identified from which articles were selected over a nine-year period. We chose to cover the period between 2007 and 2015, as few research articles had been uploaded in an exploitable PDF format prior to this.

Tables 1 and 2 summarise the articles distributions according to their source and year of publication.

Journal	Journal acronym	Number of articles (in the research corpus)
EPMA Journal – not specific to rare diseases	EMPA	4
F1000 Research – Rare Diseases channel	FR	47
Intractable & Rare Diseases Research	IRDR	13
Journal of Rare Cardiovascular Diseases	JRCD	34
Journal of Rare Disorders – Diagnosis & Therapy	JRDDT	7
Orphanet Journal of Rare Diseases	OJRD	357
Rare Diseases	RD	36
Rare Diseases and Orphan Drugs	RDOD	7
The Journal of Rare Disorders	TJRD	13
TOTAL		**518**

TABLE 1. Sources of medical articles in the research corpus

Year	Number of articles
2007	45
2008	29
2009	29
2010	15
2011	37
2012	21
2013	205
2014	77
2015	60
TOTAL	**518**

TABLE 2. Articles distribution depending on the publication year

The fact that the year 2013 is over-represented is not an issue for the neological study. The aim of this "preliminary" corpus is only to identify a sample of neologisms that appeared during this nine-year period.

3 Neological identification and extraction

3.1 Markers detection strategy

To identify and extract a sample of rare-disease neologisms from the preliminary corpus it is paramount to identify the notion of *neologism*. Any neological study presupposes knowing from when and until when a word or term is new, and sometimes even identifying when a word or term was first pronounced or written (Estopà: 2009).

For that matter, it is judicious to resort to a methodology inspired by Pearsons' use of *linguistic signals* (1998), Rey's perceived *feeling of novelty* (1972), Cabré's *neologicity criteria* (1998), and Cabré and Estopà's *neologicity filters* (2009).

Pearson (1998: 130) introduced the notion of *linguistic signals* – a list of words or expressions such as: *termed*, *named*, "*...*", *the denomination*, etc. occurring next to potential terms (or term-candidates). These allow to extract a list of those term-candidates from a text corpus. The *linguistic signals* may also be used or adapted to identify recently created terms. Time adverbs such as "recently" or "lately" may be added to increase chances of identifying new terms.

Rey developed the concept of *feeling of novelty* which a reader is expected to experience when facing a new term for the first time. Cabré defines this *feeling of novelty* as "*a psychological or socio-psychological criteria*" which is expected to be experienced within a specialised community. In order to address this *feeling of novelty* the author is thus very likely to introduce new terms alongside definitions, inverted commas and/or a gloss on the new term. Therefore, verifying the presence of neologism markers next to potential neologisms is a sensible filter to be implemented. Queries may be performed in the corpus, using terminological and neological markers as keywords.

Markers, suggested by Pearsons, Cabré and Estopà, have been divided below into categories (table 3: denomination, table 4: time, table 5: author's comments, table 6: definition, table 7: terminological variation). These lists were further enriched with words found in our research corpus, and which also introduced valuable information regarding the context of the neologism.

"..."
The term
The name
The denomination
Under the name
Mentioned
A new concept
Termed / named / called / considered
The typical

TABLE 3. Markers introducing a new denomination

Previously...	
Recently...	...named, termed, referred to (as), known (as), applied to (noun applied to), agreed, called...
Formerly...	
Originally,	
Initially...	
First reported as	
Preterit use	***(The authors proposed)***

TABLE 4. Time markers

Time markers often introduce terms that have fallen into disuse when they are followed by denominational markers such as *named*, *termed*, *called*, etc. However, when concepts become obsolete, for example in the situation of an outdated diagnostic technique being replaced by a more appropriate one, there is no new terminological creation. The term that used to name the old diagnostic technique becomes obsolete. The terminological neologism that refers to the new diagnostic technique is a primary creation.

(as) to deserve the term / the name
Should be renamed / should be preferred
We proposed XXX be renamed
It should be referred to as a proper name
So called
Considered as
The proposed term
The use of the word / term

TABLE 5. Markers bearing a comment from the author

Markers bearing a comment are also particularly interesting as they provide a unique snapshot of either the author's intellectual neological creation process or – should the neologism not have been created by the author – the author's opinion on the acceptability of the neologism. Concepts are also sometimes renamed by researchers who consider their own denomination more appropriate.

Defined as
Described as

TABLE 6. Markers introducing a definition

Neological creation may also originate from terminological variation. Here are some examples of introductory markers for terminological variation.

Also called / named / known (as)
Can also be termed
Called by some authors
Often labeled as

TABLE 7. Markers introducing terminological variation

In our study, after performing various in-corpus keyword queries, using the various markers likely to accompany neological creations, 247 character chains have been found (mostly polylexical). There is, at this stage, no absolute guarantee that these 247 character chains all appeared from 2007 on, nor that they are necessarily medical terms. In order to address this issue, we need to proceed to verifying their terminological and neological statuses.

3.2 Terminological status verification

When most of the potential neologisms identified in the corpus are polylexical units, as was the case in our research, it is paramount to ensure that they are terms referring to a specific concept and not a collocation or an explanatory or definitional paraphrase of said concept.

In order to illustrate this point, we may compare two polylexical units extracted from the corpus:

- *Genital Renal Ear Syndrome* was detected using the neologism marker *the term*. It is a very rare syndrome characterized by malformations of the kidneys, the genital tract and the middle ear. *Genital Renal Ear Syndrome* refers to a medical concept (a syndrome). Its terminological status is validated.
- *Episodic protracted vomiting attacks* has been identified using the marker *termed* located nearby. However, the co-text shows that it is a definitional paraphrase used to describe a symptom. It is followed by the marker *termed*, which introduces the actual term referring to this symptom, *dysautonomic crises*.

Sometimes, the distinction between true terms and definitional collocations or descriptive periphrases can be less clear-cut. This often happens with neologisms that are created on the spot to solve a terminological gap, as this excerpt suggests:

> *In adulthood, a more generalised symptomatic severe polyneuropathy occurs in about 3-5% of patients, often associated with an* ***"onion bulb" appearance*** *on nerve biopsy.* (Evans, 2009)

The indefinite determinant *an* is characteristic of a periphrasis. If the polylexical unit had been a term, the author would have probably inserted a definite determinant. However, a definite determinant would also suggest that the reader is already familiar with this polylexical unit. But, if the author wanted to create and introduce a new term for the first time, it would most likely have been accompanied by an introductory gloss such as: '*…associated with what we suggest to name "onion bulb" appearance*'. Moreover, the fact that only part of the expression '"*onion bulb*" *appearance*' is in quotation marks should be enough to alert us about its non-terminological status.

A few years later, in 2014, another article was published in the same journal:

> *Nerve biopsies show decreased density of myelinated nerve fibres, most pronounced in biopsies taken in the first year of life. The mean g-ratio (axon diameter versus fibre diameter) is significantly lower than normal [63].* Characteristic ***onion bulb formation*** *occurs after the age of six.* (van Paassen et al., 2014).

The idea of comparing the shape of a protrusion on a nerve to that of an onion bulb is reused, five years later, by different authors, for two rare diseases affecting the nervous system: neurofibromatosis and neuropathy. The metaphor is reused, however a variation is introduced when replacing the word *appearance* with *formation*.

In the second excerpt, the adjective *characteristic* suggests that the analogy of the onion bulb has already been used a number of times so that the authors consider that the reader already knows this *characteristic* trait. Here, the terminological and neological statuses of onion bulb formation are not clear-cut, whereas in the first excerpt, *onion bulb appearance* is more like a descriptive periphrasis. The co-text extracted from the research corpus does not yet make it possible to validate this syntagm as a term accepted by the medical community.

At this stage, a query should be made using "*onion bulb appearance*" as a keyword on *Google* and *Google Scholar*. If this polylexical unit occurs again and refers to the same concept (here a symptom) in other medical publications, it is very likely to be a term.

Particular attention has to be paid when differentiating a term from a descriptive periphrasis. Descriptive periphrasis may give rise to a term. The above-mentioned excerpt may be seen as a first attempt to formalise a neologism. Such phenomenon is called *terminologisation*. It may be observed in diachronic corpora.

Our study in co-text has proven the terminological status of the 31 following chains of characters presented in table 8.

Medical neologisms	*Year of 1st occurrence*	*Source of 1st occurrence*	
Affymetrix Genotyping Console	2007	Affymetrix Inc (2007) Affymetrix Genotyping Console 2.0 – User Manual. Affymetrix Inc.	1
aggressive vascular abnormalities of bone	2010	Orphanet Journal of Rare Diseases	2
altered immunoreactivity of pituitary polypeptide 7B2	2015	Journal of Rare Disorders	3
ARX pen holding	2014	Orphanet Journal of Rare Diseases	4
ARX-related apraxia	2014	Orphanet Journal of Rare Diseases	5
Autosomal Recessive Cerebral Atrophy	2013	Orphanet Journal of Rare Diseases	6
biotinylated proteolytic derivative of PFO	2014	Orphanet Journal of Rare Diseases	7
catalase-lacking peroxisomes	2013	Orphanet Journal of Rare Diseases	8
Cathepsin K-expressing Chondroid Progenitors	2013	Rare Diseases	9
Congenital Cockayne Syndrome	2010	Neuroradiology: The Requisites (Livre)	10
disseminated cystic bone angiomatosis	2010	Orphanet Journal of Rare Diseases	11
EBS generalized other	2008	Journal of the American Academy of Dermatology	12
Genital Renal Ear Syndrome	2007	Orphanet Journal of Rare Diseases	13
germline DICER1 truncating mutations	2015	F1000Research	14
ground glass cornea appearance	2008	Orphanet Journal of Rare Diseases	15
Hereditary sensory neuropathy type IB	2008	Orphanet Journal of Rare Diseases	16
HGSC Mercury analysis pipeline	2013	Genome Medicine	17
ION Torrent Personal Genome Machine	2015	F1000Research	18
Joint Neuromuscular Biobanks	2013	Orphanet Journal of Rare Diseases	19
LMNA-linked lipodystrophy	2007	The Journal of Clinical Endocrinology & Metabolism	20
lysosomal cysteine cathepsin K	2011	Orphanet Journal of Rare Diseases	21
MAGEL2 loss of function	2013	Nature Genetics	22
mosaic DICER1 hotspot mutations	2015	F1000Research	23
Multicore myopathy with external ophthalmoplegia	2007	Orphanet Journal of Rare Diseases	24
nCounter Digital Analyzer	2009	NanoString Technologies, Inc. (Company document)	25
nCounter Prep Station	2009	NanoString Technologies, Inc. (Company document)	26
Neomorphic RNase IIIb domain function	2015	F1000Research	27
neuro-cardio-facial-cutaneous syndrome	2008	Orphanet Journal of Rare Diseases	28
non specific XLID	2011	American Journal of Medical Genetics Part A	29
pile d'assiettes profile	2007	Orphanet Journal of Rare Diseases	30
punchinello aspect	2007	Orphanet Journal of Rare Diseases	31

TABLE 8. List of medical neologisms extracted from the research corpus

3.3 Neological status verification

The proximity of the different markers does not guarantee the neological status of the term they accompany. If we define as neologism any medical term that appeared from 2007 onwards, the date of first appearance needs to be checked for each potential neologism.

Many medical terms can be found on websites, forums and blogs aimed to a general audience, to families of patients, and managed by healthcare professionals, by patients themselves or by their own relatives. It is therefore necessary to also take into account occurrences on websites that are intended for non-specialists on a public search engine.

The main limitation of data collection via *Google* is that the first dates of publication or posting on the websites, forums and blogs that contain the studied neologisms, are not documented. A solution could be to retrieve the dates of the last updates for these webpages. This data is accessible on their cached version. When not knowing exactly when the neologism was first used by the authors of these webpages, it is possible, on the other hand, to check when the pages were updated for the last time and therefore when the use of the neologism was still considered acceptable. A neologism thriving and subsequently integrating a specialised language also depends on its *degree of acceptability* by the reader. A high *degree of acceptability* will lead them to approve of the use of the term in a specific context and to reuse it in their own writings. It may be assumed that the *degree of acceptability* for a neologism may be reflected by the number of times it is reused in the specialised literature. It is highly likely that if the author of a page no longer considers the use of a neologism as acceptable, they will correct their writings at the time of the update and will erase or replace it by another term considered more acceptable.

The *degree of acceptability* is also what guarantees the longevity of a word or term. Those that fall into disuse and eventually disappear are those whose use is, progressively, no longer considered acceptable.

While using *Google*, another question arose: if *Google* does not allow to know the first date of use, how to be sure that they did not appear before 2007. This question, though relevant, does not raise any particular issues since it is very improbable that the first occurrence of any specialised term should occur in popularised literature such as forums, blogs and websites aimed to the general public. Since the scientific researchers are those responsible for naming new concepts, it is fair to assume that the first occurrences of specialised neologisms are to appear in research articles.

4. Corpus expansion

4.1 Limitations of the preliminary corpus

In our research, most of the identified neologisms appear only once in the corpus, whereas they occur multiple times in the *Pubmed* and *ScienceDirect* databases and in the *Google* and *Google Scholar* search engines. Our "preliminary" corpus is relatively small and does not allow for the collection of a representative and reliable amount of quantitative and qualitative data.

Since it is impossible to include the whole literature related to any specialised subject, it is, then, always, necessary to integrate further specialised literature, likely to contain occurrences of the studied neologisms and variants. In order to obtain reliable data by collecting use frequencies, distribution rates, co-text and context information, etc., it is necessary to integrate most publications containing at least one occurrence of the neologism or its terminological variants, regardless of the publication date. It is important to consider that variant creation may predate the birth of the studied neologism.

4.2 Types of additional publications

In order to gather a variety of sources, it is sensible to select new texts from research articles, doctoral theses, *Powerpoint* presentations, medical books, but also from *Google* webpages intended for the general public – and sometimes created by non-experts such as patients or families of patients.

The type of source is an important piece of information in that the degree of dissemination of a neologism is to be considered as a *survival factor*. It does influence the other two survival factors, namely the neologism's use frequency and distribution rate. The more disseminated a neologism is, and therefore the better known it is by the entire medical community stricto sensu (physicians and medical researchers) and lato sensu (associations and families of patients), the more likely it is to be reused in a greater number of publications. The fact that a neologism also appears on webpages intended for the general public, which are sometimes written by non-specialists (patients, families, etc.), shows evidence of a rather high degree of dissemination.

It is also relevant to specify the degree of specialization of the publication in which the neologism appears – be it theses, research articles, papers for congresses, medical courses created by professors. In our research, it was relevant to integrate websites of associations fighting against a rare disease and providing information for patient's families, forums or blogs run by non-specialists (generally suffering from a rare disease).

In the case of a term that has fallen into disuse, one could consider the possibility that the website creator may change the terminology. This assumption should

be considered with great caution, since non-specialised authors' proper use of terminology may vary from one individual to another. Nevertheless, the use of terminology is only partially prescriptive – in fact, the standards required to harmonize medical terminology are better met in texts with a high degree of specialisation (often including authors' comments on terminological relevance) than in popularised texts. However, it may be argued that the fact that some terms remain on some popularised websites may help contribute to their survival, at least among a non-specialised community. The number of visits on these webpages is also most probably contributing to the survival and spread of the term, although further research would be needed to confirm this idea.

5 Results: corpus exploitation

5.1 Quantitative data to be retrieved

To obtain a neologism's use frequency, it is necessary to look at each publication and, use the search function, to locate and count the total number of occurrences. When access to the whole article is denied, it is still very interesting and useful to take into account the part that remains visible to all (titles, abstracts or even overviews of the articles in the list of results generated by a *Google Scholar* search).

Furthermore, the more publications a neologism appears in, the greater its chances of becoming part of the medical language, since it increases its visibility among the scientific community. Likewise, the more authors reuse a neologism, the greater also its chances of becoming part of the medical language.

5.2 Qualitative data to be retrieved

Three different types of qualitative data relating to neologisms and terminological variants have been analysed: co-text of occurrence, location of occurrence and reference articles which are particularly focused on the concept the term refers to.

Co-text around neologisms and terminological variants contain a wealth of information, such as descriptions, comments, and opinions, to track their evolution from first occurrence until potential integration in the specialised language.

For instance, table 9 shows a new pathology that needs to be named as early as 2002, but the terminological status of the first reference to it, *Hereditary sensory neuropathy with gastroesophageal reflux induced cough* is not absolutely clear-cut. This is most probably still a syntagm.

In 2003, the name becomes simpler (the information that one of the two additional symptoms is induced by the other disappears from the name). A formal stabilisation of the name *Hereditary Sensory Neuropathy with Cough and Gastroesophageal Reflux* is observed. However, this formal stabilisation of the first name goes hand in hand with the appearance of variants (*HSN I with cough and GER*, *HSAN IB*). These variants, built from the first denomination, partially acronymise the latter, probably considered too long by the authors. The underlined co-text shows that the concept has recently been identified and presents the researchers behind this discovery. Three of the authors who wrote the 2003 article also wrote the first article published in 2002, in which the syndrome is mentioned for the first time. According to the co-text, two of them are responsible for the discovery of this syndrome (P. J. Spring, J. D. Pollard). We can consider that the authorship of the first denominations belongs to them as well as to their colleagues who participated in the writing of the first two articles. Further and later co-texts have been identified, tracing the evolution of the term and of its variants.

The location of the neologism and of its terminological variants is also paramount to the degree read-

2002_Journal of the Neurological SciencesA_1 G.A. Nicholson, C. Kok, M. Kennersen, P.J. Spring, A. Ing, J.D. Pollard.	Linkage Studies in Autosomal Dominant **Hereditary Sensory Neuropathy with Gastro-Oesophageal Reflux-Induced Cough** *(Title)* Objective: To map the chromosomal location of the mutation causing dominant **hereditary sensory neuropathy (HSN) with gastro-oesophageal reflux (GER) induced cough.**
2003_American Journal of Human GeneticsA_1 C. Kok, M. L. Kennerson, P. J. Spring, A. J. Ing, J. D. Pollard and G. A. Nicholson	A Locus for **Hereditary Sensory Neuropathy with Cough and Gastroesophageal Reflux** on Chromosome 3p22-p24 *(Title)* **HSN I with cough and GER** was recently identified by two authors of the present report (J.D. Pollard and P.J. Spring), and detailed clinical and neurophysiological studies of this family have been described elsewhere (Spring et al. 2002).

TABLE 9. Qualitative data for "Hereditary Sensory Neuropathy with Cough and Gastroesophageal Reflux"

ers may be exposed to them. Qualitative analysis of the neologisms in the research sample, as well as of their variants and equivalents, attests that their location in the publications is a reliable indicator for its integration into the medical language. Indeed, the aim here is to identify which, among neologisms and variants, benefit from the most "visible" locations in the publications where they appear (main title, section titles, keywords, abstract, introduction and conclusion) as opposed to less "visible" locations (body of text, figures, and legends).

The last qualitative factor consists in identifying the reference articles contributing to the dissemination of the neologism or its variants. This could be done by reviewing the bibliography and identifying the referenced articles containing the neologisms or one of its variants in their titles. This aspect of the qualitative analysis borrows certain principles from the analysis of the location of the term in the source and the study of the terminological co-text when the latter includes references to authors and articles that are key to the terms studied. The aim here is to identify the number of times a specific article is cited in the literature published subsequently.

The same study could be carried out regarding reference authors who are considered experts on a particular concept (such as those referring to diseases) by their peers.

5.3 Corpus usability

When performing queries in the corpus, it is fundamental to be able to tell easily and efficiently, for each neologism or variant, the following information: publication date, scientific journal (if applicable) and type of publication.

Therefore, each publication was labelled, based on these associated metadata. An acronym may be assigned to each journal name and another one to each publication in order to integrate them into the naming system. Likewise any further publication added on the corpus during the corpus expansion phase should be renamed accordingly. An incremented index at the end of the new name allowed differentiating publications already sharing all other metadata. The resulting filenames were of the form YEAR_TYPE_JOURNAL_INDEX.

Using the IMS-CWB concordancer, along with this labelling system, allows to quickly find the year, the journal, or the type of publication without having to open the corresponding file. Since the name of the original file appears next to each line of occurrences of the search term in IMS-CWB, it is then easy to obtain all the metadata without having to go and search for it manually.

6 Discussion

The presented methodology enables to create a diachronic research corpus targeted on identifying and thoroughly monitoring the life-cycle of a sample of terms from their first occurrence until today or until their last occurrence (should the neologism or term fall into disuse).

Corpus-building methodologies for neological study already exist such as those developed by the *Observatori de Neologia* (University Pompeu Fabra in Spain) or *Néoveille* (Sorbonne-Paris-Cité, EMPNEO and the University São Paulo in Brazil), however the novelty here lies in the possibility to gather extensive data to observe the evolution of any neologism or term at any stage during its life-cycle. This could also be applied to defining periphrasis undergoing a terminologisation process in order to better understand the mechanisms at work when a periphrasis becomes a term.

The examples given herein represent a subset of the dataset used in our doctoral thesis (Schneider, 2020) and enabled to acquire more information and a better understanding of the mechanisms at work behind terminological creation as a well-thought out act by the speakers of a linguistic community. Similarly, it also allows to better grasp the underlying mechanisms behind terminological phenomena such as semantic shift, terminological variation and resemantisation.

Limited access to online journals and exclusion of oral medical contributions (such as conference recordings) is currently the main limitation in this methodology. The thorough study of neological survival criteria is based on a significant amount of data rather than on an exhaustive amount of data. Adapting this methodology to integrate oral scientific literature would be particularly relevant considering the importance of congresses, seminars, and academic lectures in the field of rare diseases for instance.

7 Conclusion

The methodology presented here has been developed and designed for the study of neological creation mechanisms and the possible reasons why some terms may "survive" and integrate a specialised language or fall into disuse. It allows the study of new terms in context, from the date of their creation until today. The *survival factors* are relevant criteria to assess the degree of integration of terms, regardless of the year when they first occurred and of the domain they belong to. This list of quantitative and qualitative factors may also be further enriched in the future as other criteria likely to contribute to the integration of a new term, may be identified.

In conclusion, the essential steps of the methodology are: (1) Identify and monitor the life cycle of a sample of specialised neologisms and characteris-

tics including years and languages of the neological corpus. (2) Defining the specialised field under study to select literature specific to that field and whether to integrate popularized literature. (3) Identify neologisms by resorting to neological, time, gloss, definition, and variation markers and verify the terminological and neological statuses of all character chains. (4) Expand the corpus which may not be representative enough of the whole medical literature, to gather sufficiently reliable quantitative data such as use frequency and distribution rate. (5) Collecting the necessary data according to the research objectives.

This new methodology allowed for the quantitative analysis, as part of our doctoral thesis, of each of these terms and the comparison of the temporal evolution of use frequency and distribution rate of each neologism and its variants.

The study in co-text, the location and the reference articles allow to refine the conclusions drawn from the study of the quantitative data, while nuancing the degree of "visibility" of each occurrence, depending on its location in publications and on the number of reference articles dedicated to the concept to which the neologism or variant refers. The co-text analysis offers a hindsight of the authors' motivation and opinion around the creation and use of any given neologisms or variants.

This data helps detecting which variants are the most frequently used, by the greatest number of articles, over the years in order to try to predict those that would integrate a specialised language on a long-term basis.

Further research could be carried out to demonstrate the influence of authors as part of *survival factors*. Distribution rate could be studied in terms of numbers of authors reusing the term instead of numbers of publications in which a term appears. Similarly, reference authors who are considered experts in the concept to which the neologism or variant refers, could be further investigated to measure the degree of visibility they may provide to said neologism or variant.

Bibliography

AUGER, Pierre, 2010. Pour des critères extralinguistiques de néologicité. In Actes del I Congrès Internacional de Neologia de les Llengües Romàniques / coord. par Maria Teresa Cabré, Ona Domènech, Rosa Estopà, Judit Freixa and Mercè Lorente, pp. 117-121.

CABRÉ, Maria Teresa, 1998. Terminologie : théorie, méthode et applications. Les presses de l'Université d'Ottawa, Armand Colin, 322 p.

CABRÉ, Maria Teresa, 2006. La clasificación de neologismos: una tarea compleja, Alfa, Sao Paulo, Vol. 50, n°2, pp. 229-250.

CABRÉ, Maria Teresa, Rosa, ESTOPÀ, 2009. Les paraules noves: criteris per detectar i mesurar els neologismes. Vic; Barcelona: Eumo; Universitat Pompeu Fabra.

ESTOPÀ, Rosa, 2009. Neologismes i filtres de neologicitat: aspectes metodològics. In: Cabré, M. Teresa; Estopà, Rosa (ed.). Les paraules noves: criteris per detectar i mesurar neologismes. 1 ed. Barcelona: Universitat Pompeu Fabra. Pp. 41-48.

ESTOPÀ, Rosa, 2015. Sobre neologismos y neologicidad: reflexiones teóricas con repercusiones metodológicas. In Alves, I. M.; Simões Pereira, E. (eds.). Neologia das Línguas Românicas. São Paulo: CAPES; Humanitas. Pp. 111-150.

EVANS, Gareth R., 2009. Neurofibromatosis type 2 (NF2): A clinical and molecular review, Orphanet Journal of Rare Diseases, 4: 16.

FREIXA, Judit, 2010. Paraules amb rareses. Terminàlia, 1 : 7-16.

HUMBLEY, John, 2003. (Chapitre d'ouvrage), La néologie en terminologie, Jean-François Sablayrolles. L'innovation lexicale, Champion, pp.261-278.

INSERM, 1997. Orphanet: an online database of rare diseases and orphan drugs. Copyright. Available at http://www.orpha.net Accessed 2016.

PEARSON, Jennifer, 1998. Terms in Context. John Benjamins Publishing Co., Amsterdam, 243 p.

QUIRION, Alain, 2010. Pour des critères extralinguistiques de néologicité. In Actes del I Congrès Internacional de Neologia de les Llengües Romàniques / coord. par Maria Teresa Cabré, Ona Domènech, Rosa Estopà, Judit Freixa and Mercè Lorente, pp. 123-127.

RENOUF, Antoinette, 2007. Corpus development 25 years on: from super-corpus to cyber-corpus. In Corpus Linguistics 25 years on, ed. R. Facchinetti, 27–49. Amsterdam/New-York: Rodopi.

Rey, Alain, 1976. Néologisme : un pseudo-concept ?, Les Cahiers de Lexicologie, n° 28, pp. 3-17.
Sablayrolles, Jean-François, 1996-1997. Néologismes : une typologie des typologies. Cahier du CIEL, UFR EILA, Université Paris 7, pp. 11-48.
Sablayrolles, Jean-François, 2000. La néologie en français contemporain. Examen du concept et analyse de productions néologiques récentes, Paris, Honoré. Champion editor, coll. « Lexica », n° 4.
Sager, Juan Carlos, 1989. Term Formation, Lebende Sprachen, n°34, 159-161.
Schneider, Coralie, 2018. Determining survival probabilities for specialised neologisms in medical English and French: a diachronic perspective, ASp, n° 74, 53-76.
Schneider, Coralie, 2020. (PhD) Identification de facteurs d'intégration quantitatifs et qualitatifs des néologismes à la langue spécialisée médicale : une analyse sur corpus diachronique entre 2007 et 2015. Université Paris Diderot (France), Universitat Pompeu Fabra (Spain).
van Paassen, Barbara W. et al., 2014. PMP22 related neuropathies: Charcot-Marie-Tooth disease type 1A and Hereditary Neuropathy with liability to Pressure Palsies, Orphanet Journal of Rare Diseases, 9: 38.