A contribution of discourse analysis to the morphology of nominalizations. Study of the use of nominalizations in the genre of the scientific activity report using a corpus linguistics approach based on the Démonette and Lexique 3 databases.

Hugo Dumoulin
MoDyCo, Université Paris Nanterre

The purpose of this paper is to explore the parameter of discourse genre in the study of nominalizations. How does discourse genre constrain the way nominalizations are used in discourse, from a morphological, syntactic and semantic point of view? We will look at these different aspects using a linguistic approach based on a corpus of professional writings in the activity report genre. Based on the work of Née, Sitri, Véniard (2016), who propose to make the link between genres and discursive routines, we propose to highlight the role of nominalizations in the specific phraseology of the activity report. In turn, it will be shown that discourse genre has affinities with certain types of verbal nominalizations, which helps to bring the discourse parameter to the fore in order to distinguish the competing suffix derivations -ion, -ment or -age, following the perspective developed by Missud & Villoing 2020.

1 Corpus and methodology

1.1 Corpus

The RapportS corpus was compiled as part of the ArchivU[1] project, which aims to approach institutional discourse from the angle of professional writings. It brings together the HCERES 2018 self-evaluation reports of 38 laboratories at the University of Paris Nanterre, which have been occluded and encoded in XML format[2]. To highlight the specific behaviour of nominalizations as a function of the discourse genre parameter, we rely on several comparison corpora: first of all, the Scientext corpus, which brings together texts of scientific articles (Tutin & Grossman 2014), secondly, the ADMIN corpus consisting of activity reports from two (non-scientific) French administrations (ADMR and Ucanss) in 2018, then several corpora from various discourse genre : the VŒUX corpus (Leblanc 2016), gathering the annual greetings of the presidents of the French republic since the 1970s, and the two corpora of the Lexique 3 database (New, Pallier, Brysbaert & Ferrand 2004), namely a written corpus composed of novels from the 1950-2000 period taken from Frantext[3], and an corpus of film subtitles. The aim is to provide means of gaining a detailed understanding of the linguistic constraint exerted by the discourse genre, by drawing a comparison between genres that are far apart (argumentation/fiction) or more closely related by their theme (scientific article/scientific text) or function (activity report in the science field/in other fields).

1.2 Methodology

Our tool-based linguistic approach is rooted in the general field of textometry (Lebart & Salem 1994). The TXM software (Heiden, Magué, Pincemin, 2010) is used for lemmatisation and automatic syntactic annotation (TreeTagger) of our corpus as a textual database. We chose to annotate very precisely the deverbal nominalizations in Xment, Xion, Xage of our corpora by using the Démonette database (Hatout & Namer 2014) via a Python script of our own. As a

[1] Labex *Les passés dans le présent.*

[2] The corpus represents 921,989 occurrences for 40,034 forms according to the segmentation and indexing performed by TXM.

[3] ATILF. *Base textuelle Frantext* (En ligne). ATILF-CNRS & Université de Lorraine. 1998-2023 https://www.frantext.fr/

result, we carry out a set of lexicometric statistical observations, but also textometric ones, such as the calculation of specific co-occurrents.

2 Results

2.1. The distribution of nominalizations in RapportS compared with other corpora

2.1.1. The contribution of verbal nominalizations in characterizing the genre of the report

We obtained initial lexicometric results (Figure 1): the frequency of deverbal nominalization tokens appears relatively high (3.40%) in RapportS and in ADMIN (2,93%) compared with VOEUX (0.96%), or even with LEXIQUE3 (0.36%). In addition, the type/token ratio appears very low in our corpus (0.058) compared with the VOEUX corpus (0.31), indicating a stronger fixation of vocabulary in the reports. In the light of these frequency calculations, verbal nominalizations appear to be somewhat characteristic of the genre of the professional activity report. Nevertheless, it appears that the frequency of deverbal nominalizations is also very high in the corpus SCIENTEXT of scientific texts (3,59%).

	Corpus size	Verbal nominalizations			
		Type	Token	Tokens frequency (%)	type/token ratio
RapportS	921 898	1 817	31 328	3,3982	0,0580
VŒUX	118 719	362	1139	0,9594	0,3178
SCIENTEXT	3 320 474		119396	3,5958	
ADMIN	43 145		1266	2,9343	
Lexique3_romans	~14 700 000			0,36089	
Lexique3_films	~50 000 000			0,20973	

Figure 1. Distribution of verbal nominalization among discourse genres

2.1.2. An attraction of the -ion derivation for the discourse genre of the activity report and of the scientific article

A factorial correspondence analysis (Benzécri 1973) was carried out to represent the distribution of derivational suffixes among the corpora. The relationship between the modality of suffix and the modality of discourse genre appears to be statistically significant, although of low intensity ($\phi = 0.1168$). It appears that the axis of greatest inertia opposes the suffix -ion to the other suffixes -ment and -age (87.83% of the variance), while the secondary axis opposes the suffixes -ment and -age to each other. Under these conditions, Figure 2 clearly shows a statistical attraction of the suffix -ion for the report (RapportS, Admin) and the scientific article (Scientext) genres, while -ment is significantly attracted by the Vœux and Lexique3_romans corpora. Finally, -age appears to have a positive association with the Lexique3-films corpus.
The more concrete nature of -age (Missud & Villoing 2021) seems to be confirmed by its attraction to the subtitle genre, representing oral interaction in written form. Conversely, we can link the preference for -ion in activity reports and scientific articles to the greater abstraction that characterizes it compared to -ment (Martin 2008: 165). Nevertheless, the -ion suffix does not clearly show a preference for the genre of the scientific article or the genre of the activity report.

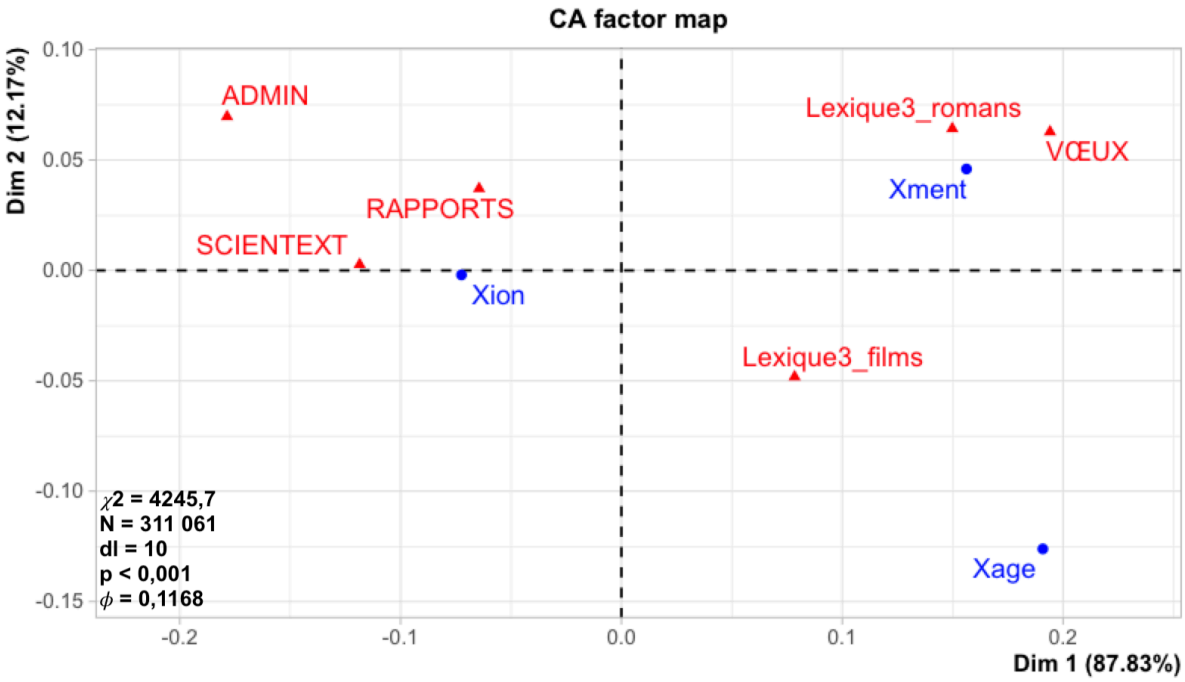

Figure 2. Correspondence factor analysis representing the distribution of derivationnal suffixes among the corpora

2.2. Textometric study of the corpus : nominalization as part of a routinization process

The textometric study now makes it possible to study the behaviour of nominalizations in the RapportS corpus by restoring them in their phrasal and textual context. In particular, we seek to identify the specific prepositional complements of nominalizations in the corpus, using Corpus Query Language (Figure 3). It appears that, among the statistically specific co-occurrents (Lafon 1980), we find a large number of lexemes that are part of the lexical field of scientific activity: unit, research, knowledge, data, product, tool, ecosystem, knowledge, etc. It should be noted that the specific co-occurrents of -ion ('unit', 'knowledge') are considerably more abstract than those of ment ('tool', 'researcher'). But generally speaking, the use of nominalizations in the corpus is specifically linked to the representation of scientific activity - a central theme which carries the very function of the genre: we are dealing with a discursive routine characteristic of the genre which links deverbal nominalizations and the lexical field of scientific activity.

Requête [demonext="Xion"][frlemma="de|du"]

Cooccurrent	Fréquence	CoFréquence	Indice	Distance moyenne
NOM unité	1912	254	54	3.2
NOM produit	260	84	48	1.2
NOM connaissance	337	82	36	1.4
VER:pres rechercher	56	35	32	3.0
NOM savoir	332	76	32	2.0
NOM donnée	709	111	30	1.9
NOM écosystème	82	39	30	2.0
NOM recherche	5080	403	29	3.0
NAM Introduction	28	23	26	3.0
NOM colloque	1067	123	21	2.3
NOM activité	1364	143	20	3.3

Requête [demonext="Xment"][frlemma="de|du"]

Cooccurrent	Fréquence	CoFréquence	Indice	Distance moyenne
NOM recherche	5080	157	21	2.9
NOM outil	325	24	10	2.2
NOM chercheur	1200	48	10	2.3
NOM donnée	709	35	10	1.9
ADJ nouveau	1440	53	10	2.5
NOM intérêt	214	19	10	1.6
NOM fond	41	10	9	1.0
DET:ART un	13124	253	8	2.9
NAM CTEM	8	5	7	3.2
NOM signal	15	6	7	1.0
NOM doctorants	922	35	7	2.4
NOM professeur	351	20	7	2.5

Figure 3. Specific co-occurrents in the prepositional position of nominalizations in -ion (left) and -ment (right). Distance to the right : 5 occurrences.

The importance of this result for the discourse semantics of nominalizations becomes apparent if we relate it to their syntactic properties. As Martin (2008) points out, -ion and -ment nominalizations are susceptible to multiple semantic subspecifications, leading to several readings: causative/inchoative, agentive/non-agentive. As a result, thematic roles that are clear for verbal complements become ambiguous for the prepositional complements of verbal nominalizations. For example, in the expression: "The development of new means of communication", there is ambiguity between the transitive and intransitive readings. Thus, the lexical field of scientific activity, which enters specifically into the position of complement of nominalizations, comes with an unclear semantic status. This contributes to the hypothesis of an abstract, routinised representation of research activity. The actants disappear, and things seem to happen by themselves - which may justify, notwithstanding a few remarks, the inclusion of nominalizations in the category of the "préconstruit " forged by the French school of discourse analysis (Pêcheux & alii 1979, Sériot 1986).

Conclusion

Thus, both because of the types favoured by the genre of the activity report (abstract suffixes in -ion), and because of their textual role as co-occurrents of the lexical field of scientific activity neutralising certain semantic values, verbal nominalizations can be associated with an effect of abstraction and discursive “routinization” specific to the activity report.

Bibliography

Benetti, L, & Corminboeuf, G.,2004, « Les nominalizations des prédicats d’action », Cahiers de linguistique française, 26, 413-435.

Benzécri, J.-P., 1973, *L’analyse des données. Tome 1. La taxonomie*, Paris, Dunod, 675 p.

Haas, P., Huyghe, R., Marin, R., 2008, « Du verbe au nom : calques et décalages aspectuels », Actes du Congrès Mondial de Linguistique Française, Paris, France, 2052-2065.

Hathout, N. & Namer, F., 2014, “Démonette, a French derivational morpho-semantic network”. Linguistic Issues in Language Technology, 11 (5), pp.125-168.

Heiden S., Magué J.-Ph., Pincemin B., 2010, « TXM : une plateforme logicielle open-source pour la textométrie – conception et développement »,10th International Conference on the Statistical Analysis of Textual Data – JADT 2010, Juin 2010, Rome, Italie, 1021-1032.

Lafon P., 1980, « Sur la variabilité de la fréquence des formes dans un corpus », Mots. Les langages du politique 1, pp. 127-165.

Lebart L., Salem A., 1994, Statistique textuelle. Dunod.

Leblanc, J.-M., 2016, Analyses lexicométriques des vœux présidentiels, Londres, ISTE éditions.

Martin, F., 2008, “The Semantic of Event Suffixes in French”, in Schäfer, F. (ed.), Working Papers of the SFB 732, vol. 1., Stuttgart, University of Stuttgart.

Missud, A., & Villoing, Fl., 2020, “The morphology of rival -ion, -age, and -ment selected verbal bases”, *Lexique*, 26, pp. 29-52.

Missud, A., & Villoing, Fl., 2021, “Investigating the distributional properties of rival -age suffixation and verb to noun conversion in French”, *Verbum*, XLIII, pp. 41-68

Née E., Oger C., Sitri F., 2017, « Le rapport : opérativité d’un genre hétérogène », Mots 114, Le rapport, entre description et recommandation, 9-24.

Née E., Sitri F., Veniard M., 2016, « Les routines, une catégorie discursive pour catégoriser les genres ? », Lidil 53, 71-93.

New, B., Pallier, Ch., Brysbaert, M. & Ferrand, L., 2004, “Lexique 2: A new French lexical database”, Behavior Research Methods, Instruments & Computers, 36, 516-524.

Pêcheux, M., Haroche, Cl., Henry, P., Poitou, J.-P., 1979, « Le rapport Mansholt : un cas d’ambiguïté idéologique », Technologies, Idéologies, Pratiques, 2, 1-83.

Sériot, P., 1986, « Langue russe et discours politique soviétique : analyse des nominalizations », Langages, 81, 11-41.

Tutin, A., & Grossman, F., 2014, *L’écrit scientifique : du lexique au discours ; autour de Scientext*, Presses universitaires de Rennes.