Pour une formalisation de la terminologie des carnets de Ugo Ferrandi : un voyage dans la culture somalienne précoloniale (Short Paper)

Silvia Piccini[1,*,†], Andrea Bellandi[1,†], Jama Musse Jama[1,2,†], and Giuliana Elizabeth Vilela Ruiz[1,†]

[1] *Istituto di Linguistica Computazionale "A. Zampolli", Via Giuseppe Moruzzi 1, Pisa,56124, Italia*
[2] *Società Geografica Italiana, Via della Navicella 12, Roma, 00184, Italia*

Abstract
The aim of this contribution is to present the initial steps in constructing a bilingual Somali-Italian termino-ontological resource, dating back to the era of Italy's colonial expansion in Africa. The terminological data were extracted from the notebooks written by the Italian explorer Ugo Ferrandi (1852-1928), published by the Società Geografica in 1903 under the title "Lugh. Emporio Commerciale sul Giuba". In order to develop this termino-ontological resource, we have adopted Semantic Web technologies (RDF, OWL, SPARQL) and the Linked Open Data paradigm, thereby ensuring the quality, accessibility, interoperability, and reusability of the data.

Résumé
L'objectif de cette contribution est de présenter les premières étapes de la construction d'une ressource termino-ontologique bilingue somali-italien, remontant à l'époque de l'expansion colonialiste de l'Italie en Afrique. Les données terminologiques ont été tirées des carnets rédigés par l'explorateur italien Ugo Ferrandi (1852-1928), publiés par la Società Geografica en 1903 sous le titre "Lugh. Emporio Commerciale sul Giuba". Afin de développer cette ressource termino-ontologique, nous avons adopté les technologies du Web sémantique (RDF, OWL, SPARQL) et le paradigme des Données liées ouvertes, assurant ainsi la qualité, l'accessibilité, l'interopérabilité et la réutilisabilité des données.

Keywords
Somali, OntoLex-Lemon, Ugo Ferrandi

1. Introduction

Cette contribution vise à présenter la construction d'une ressource termino-ontologique bilingue somali-italien remontant à l'époque de l'expansion coloniale de l'Italie en Afrique. Plus spécifiquement, les données terminologiques ont été extraites des carnets de l'explorateur

3rd International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2024, June 27-28, 2024, Granada, Spain
[*] Corresponding author.
[†] These authors contributed equally.
silvia.piccini@ilc.cnr.it (S. Piccini); andrea.bellandi@ilc.cnr.it (A. Bellandi); jama.bellandi@ilc.cnr.it (J. Musse Jama); giulianaelizabethvilelaruiz@cnr.it (G. Vilela Ruiz)
0000-0002-2584-0191 (S. Piccini); 0000-0002-1900-5616 (A. Bellandi); 0000-0002-2987-5914 (J. Musse Jam)

italien Ugo Ferrandi[2] (1852 - 1928), rédigés lors de son séjour dans le village commercial de Lugh, dans la région du Benadir.

L'œuvre de Ferrandi constitue une mine d'informations inestimables du point de vue ethnographique, anthropologique, culturel et linguistique. L'auteur décrit des objets, des concepts et des phénomènes uniques de la culture somalienne, rapportant une pléthore de termes indigènes qui représentent une phase ancienne de la langue telle qu'elle était parlée par les pasteurs et les agriculteurs nomades à l'époque précoloniale. Les carnets témoignent d'une culture passée qui était très longtemps confinée à l'oralité, le somali n'ayant reçu un système officiel d'écriture que le 21 octobre 1972, lorsqu'il a été érigé en langue officielle de la République de Somalie.

Il faut cependant souligner que les données fournies par les carnets nécessitent parfois une vérification minutieuse. D'une part, Ferrandi enregistre les mots qu'il entend sans une connaissance approfondie des dialectes locaux, en utilisant souvent des graphies incongruentes. D'autre part, les informations culturelles sont parfois imprégnées de stéréotypes et de préjugés. La vision simplifiée des premiers explorateurs a fortement contribué à la création d'une image collective où l'Afrique apparaissait comme un pays mystérieux et dangereux, habité par des tribus sauvages ayant besoin de civilisation. En stigmatisant l'Autre, l'Occident a construit sa propre identité positive [3], mettant en œuvre ce mécanisme dialectique connu sous le terme de « othering », selon le néologisme forgé par [4].

La création de cette ressource termino-ontologique s'inscrit dans la cadre d'un projet de recherche financé par l'organisation philanthropique Fondazione RUT[3], impliquant la collaboration de l'*Istituto di Linguistica Computazionale* « A. Zampolli » (ILC-CNR) et de la *Società Geografica Italiana* (SGI). L'objectif ultime de ce partenariat est d'enrichir et de rendre accessible le précieux patrimoine culturel issu des explorations du XIXe siècle en Somalie, incluant des cartes géographiques, des photographies, des artefacts et des carnets de voyage, conservés dans la bibliothèque de la SGI.

Cette ressource termino-ontologique en cours de développement sera intégrée dans l'*Osservatorio Italiano del Multilinguismo*, à savoir une base de données, constamment mise à jour et formalisée selon les technologies computationnelles avancées pour faciliter l'étude et la reconstruction des relations entretenues par le peuple italien avec d'autres cultures et langues. Les données seront rendues accessibles à travers l'infrastructure de recherche CLARIN.

Le présent article est organisé comme suit : après avoir illustré les cahiers rédigés par Ferrandi lors de son expédition en Somalie, la section 3 présente les méthodologies et les modèles utilisés pour construire la ressource terminologique. Dans la section 4, un exemple d'entrée sera présenté, et des conclusions seront tirées.

2. Les carnets de Ugo Ferrandi : un survol sur les données linguistiques et culturelles

La première phase de notre travail a consisté en l'extraction manuelle des termes somaliens et de leurs correspondants italiens à partir des carnets rédigés par Ferrandi et publiés en 1903 par la SGI sous le titre « Lugh. Emporio Commerciale sul Giuba ».

En l'état actuel, plus de 400 termes ont été collectés, couvrant un large éventail de champs sémantiques, y compris la flore, la faune, les habitations, les rites matrimoniaux et funéraires,

[2] Pour une bibliographie approfondie de Ugo Ferrandi, voir [1] et [2]. L'œuvre de Ferrandi est disponible en ligne à l'adresse suivante : https://archive.org/details/lughemporiocomm00ferrgoog.
[3] https://fondazionerut.org/

le folklore, les festivals, les vêtements, les jeux, le mobilier domestique, les armes, l'organisation sociale (Figure 1).

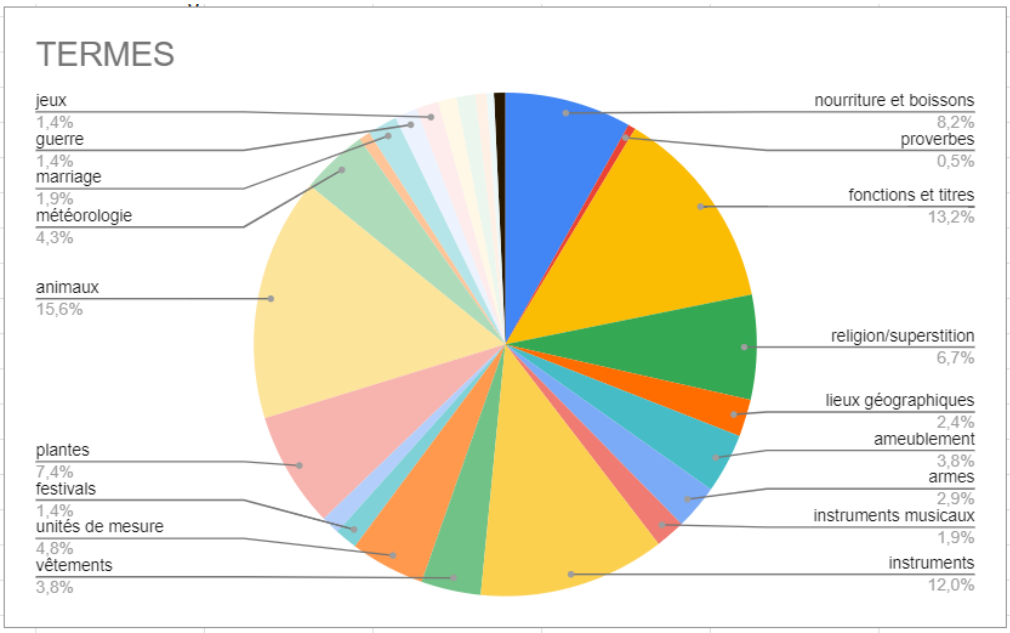

Figure 1: Les champs sémantiques des carnets de Ugo Ferrandi

Certains de ces termes figurent également dans le glossaire en annexe des carnets, comprenant 300 termes qui appartiennent, selon Ferrandi, à trois langues parlées à l'époque dans le village de Lugh, : somali (s), rahanuìn (r) et « lughiano » (l). Bien que cette catégorisation ne corresponde pas à la classification linguistique actuelle du groupe somalien, on peut raisonnablement supposer que par rahanuìn Ferrandi faisait référence au dialecte Maay, décrit dans [5] comme le somali central. Par somali l'explorateur entendait, en revanche, le dialecte généralement connu sous le nom de somali commun [6, 7], devenu ensuite la langue standard de la République de Somalie en vertu de son prestige. Ce dialecte était déjà utilisé bien avant l'avènement des puissances coloniales comme *lingua franca* pour permettre une communication plus large entre les nombreuses tribus somaliennes. Une enquête plus approfondie est nécessaire, au contraire, pour identifier le dialecte appelé par Ferrandi « lughiano ». Il est toutefois à noter la sensibilité linguistique démontrée par l'explorateur dans la collecte des données linguistique, qui perçoit immédiatement le caractère polyglotte de cette région, typique d'une époque historique où la langue n'a pas encore été fixée par une tradition littéraire, une discipline scolaire et une consécration officielle [8].

3. La ressource termino-ontologique : technologies et modèles

Afin de développer notre ressource termino-ontologique, nous avons adopté les technologies du Web sémantique (RDF, OWL, SPARQL) et le paradigme des Données liées ouvertes, assurant ainsi la qualité, l'accessibilité, l'interopérabilité et la réutilisabilité des données [9]. Le but ultime est en effet celui de publier et partager la ressource terminologique dans un Web de Données ouvert et interconnecté, et contribuer ainsi à combler l'absence du somali dans le nuage de

données linguistiques liées [10]. Le somali, en effet, langue couchitique orientale parlée par 22 millions de personnes, représente à l'état actuel une langue sous-dotée en termes d'outils automatiques et de ressources langagières. À l'exception de quelques travaux [11, 12, 13, 14], peu d'attention a été accordée jusqu'à présent à la création de ressources formalisées utilisables dans des tâches de TAL et d'Intelligence Artificielle.

Dans notre ressource, les niveaux linguistique et conceptuel sont séparés, conformément à certaines approches désormais bien établies en terminologie [15, 16, 17, 18, 19], notamment en terminologie culturelle [20][4].

3.1. Le niveau linguistique

La description linguistique des termes est confiée au modèle OntoLex-Lemon [21], qui constitue une norme *de facto* dans le domaine de la lexicographie computationnelle pour la construction de lexiques en RDF. L'architecture modulaire de OntoLex-Lemon permet une description complète et précise des caractéristiques linguistiques d'un terme et repose sur la distinction entre le plan linguistique et conceptuel de manière cohérente avec les postulats théoriques de notre travail.

Chaque entrée lexicale italienne et somalienne est définie comme une instance de la classe *Lexical Entry* et est liée à travers la propriété *ontolex:sense* à une ou plusieurs acceptions, comme dans le cas des mots polysémiques. Le sens, instance de la classe *Lexical Sense*, est défini par un ensemble de relations lexico-sémantiques exprimant les relations paradigmatiques entre les termes (hyperonyme, synonyme, synonyme approximatif, etc.). Chaque terme, en somali comme en italien, est accompagné d'une définition tirée des carnets de Ferrandi. La définition est également donnée au niveau du concept. Un exemple d'entrée lexicale sera donné dans la section 4.

La composante linguistique a été enrichie – le cas échéant – avec les données extraites du lexique intégré dans le Somali Corpus réalisé par Jama Musse Jama [11]. Ce lexique fournit des informations linguistiques précieuses, telles que la relation que le mot recherché entretient avec d'autres mots ; sa fréquence au sein de sous-corpus spécifiques ; son étymologie ; ses synonymes et antonymes ; ses variantes orthographiques ; et enfin, les définitions extraites d'une liste de dictionnaires de référence ainsi que des traductions en anglais, italien, français et suédois.

Le lexique du corpus somalien étant encodé dans un format propriétaire, la connexion avec notre ressource termino-ontologique n'a pu être établie qu'après la conversion préalable du lexique du corpus dans le modèle OntoLex-Lemon. Ce processus a impliqué une conversion des données du format propriétaire du corpus somalien au format CONLL-U (standard largement utilisé dans le cadre des Dépendances Universelles (UD) pour annoter les données au niveau de la phrase et du mot/token) et, par la suite, de ce dernier au modèle de données OntoLex-Lemon à l'aide d'un convertisseur spécialement construit à cet effet. Les termes du lexique de Ferrandi ont été associés aux termes du lexique du corpus somalien à travers la propriété *rdfs:seeAlso*. La Figure 1 montre le flux de travail du projet.

[4] L'hypothèse théorique sur laquelle repose ce travail est en effet que la terminologie est une « science double », sa spécificité résidant précisément dans la relation entre la langue et le savoir spécialisé [16, 17].

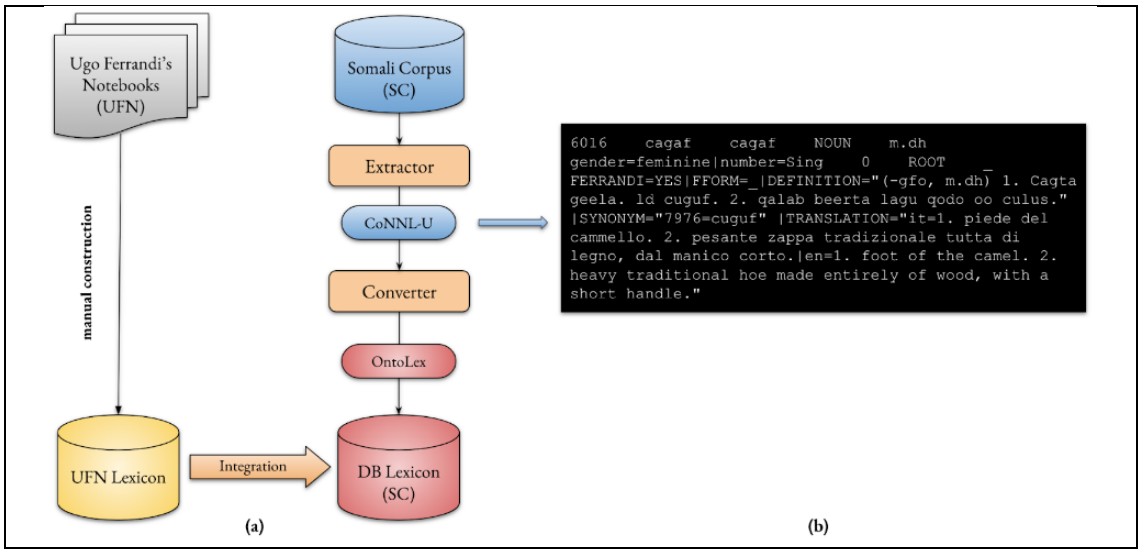

Figure 2: Le flux de travail du projet.

3.2. Le niveau conceptuel

Le sens de chaque entrée lexicale est lié à travers la relation *ontolex:reference* à un concept décrit dans une ontologie en OWL qui représente formellement la conceptualisation dominante du monde en Somalie au tournant du XXe siècle.

L'architecture de cette ontologie s'inspire largement du modèle lexical SIMPLE proposé par [22] et dont l'efficacité dans la structuration des lexiques spécialisés a déjà été démontrée [23]. Basé sur les principes fondamentaux de la théorie lexicale générative élaborée par [24], ce modèle permet de capturer effectivement la multidimensionalité des concepts et de la décrire à travers la structure Qualia. Cette dernière, avec ses quatre rôles (formel, constitutif, télique et agentif), permet d'exprimer des dimensions orthogonales du concept, allant ainsi au-delà des relations hiérarchiques de subsomption.

L'ontologie SIMPLE comprend 139 concepts organisés dans une structure hiérarchique avec 6 niveaux de profondeur. Les concepts sont interconnectés par un vaste réseau de relations, également inspirées de la structure Qualia, et organisées en relations formelles (*is-A*), constitutives (*isPartOf*, *hasAsPart*, *location*, *madeOf*, *produces*, etc.), téliques (*purpose*, *objectOfTheActivity*, *usedFor*, etc.), et agentives (*resultOf*, *causedBy*, *derivedFrom*, etc.). L'ontologie SIMPLE étant conçue pour un lexique général, ses concepts représentent les nœuds les plus élevés dans la hiérarchie et ils ont été ensuite spécialisés à travers une approche descendante pour modéliser les domaines spécifiques de Ferrandi.

4. Exploitation de la ressource

À titre d'exemple, voici la formalisation en RDF de l'entrée *barchi* « appui-tête » (Figure 3) et la formalisation ontologique du concept désigné par ce terme (Figure 4), réalisée à l'aide du logiciel Protégé.

```
barchi_entry a ontolex:Word ;
        lexinfo:partOfSpeech lexinfo:noun ;
        ontolex:canonicalForm [ ontolex:writtenRep "barchi"@som ]  ;
        ontolex:otherForm [ ontolex:writtenRep "barciuma"@som ]
        ontolex:sense :barchi_s1 ;
        rdfs:seeAlso barkin_entry.

:barchi_s1 a ontolex:LexicalSense ;
        skos:definition "Poggiatesta in legno dolce particolarmente diffuso tra i Rahanuìn
        che constava di un sostegno massiccio per le donne e di una colonna semplice o
        doppia per gli uomini. Su una base circolare o ellittica si innestava una parte
        superiore semilunare ove veniva posato il capo durante il riposo"@ita;
        ontolex:reference onto:HEADREST ;
        lexinfo:dating lexinfo:old .


barkin_entry a ontolex:Word ;
        lexinfo:partOfSpeech lexinfo:noun ;
        lexinfo:gender lexinfo:masculine ;
        ontolex:canonicalForm :barkin_cf ;
        lexinfo:geographicalVariant barkimo_entry ;
        ontolex:sense :barkin_s1 .

:barkin_s1 a ontolex:LexicalSense ;
        skos:definition "Qori si wanaagsan loo qorey, leh madax qaab bileed ah oo la
        barkado,lug, gacan gelis ama gacan qabsi ah iyo sal wareegsan, ballaaran oo uu ku
        fariisto."@som , "Poggiatesta in legno, cfr. barshin."@it , "Wooden neck-rest."@en ;
        ontolex:reference onto:HEADREST.
```

Figure 3: L'entrée lexical du lexique de Ferrandi *barchi* « appui- tête » formalisée en RDF et l'entrée lexicale correspondante du corpus somalien *barkin*.

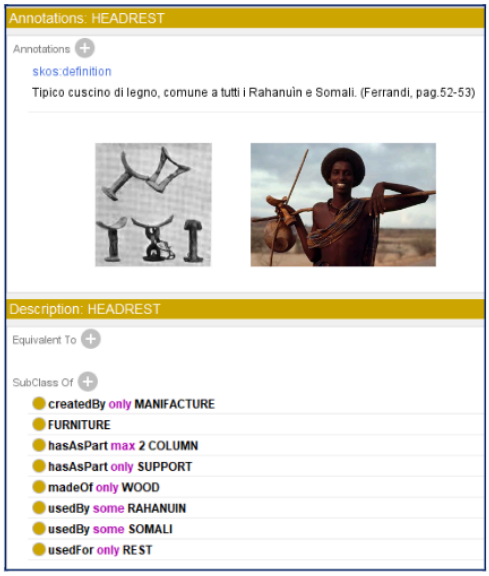

Figure 4: La formalisation du concept désigné par le terme *barkin* à travers le logicilel Protégé.

Bien que tous les termes et concepts extraits n'aient pas encore été formalisés à ce jour, il est déjà possible d'apprécier les avantages liés à une telle structuration formelle. Il est en effet possible d'effectuer des requêtes prenant en compte soit le plan linguistique, soit le plan conceptuel, soit les deux en combinaison. Par exemple, l'utilisateur peut, à travers le langage de requête SPARQL, identifier combien de termes swahilis ou arabes sont présents dans la ressource et dans quels champs sémantiques ils se concentrent davantage. À travers l'adoption de nouveaux termes, on peut discerner l'introduction d'objets inédits au sein de la culture, certains emprunts reflétant une évolution culturelle et offrant des indications sur les changements dans la vie quotidienne, ou sur la manière dont de nouveaux éléments matériels ont été incorporés et assimilés au fil du temps.

La connexion avec le lexique intégré dans le Somali corpus permet également d'apporter une profondeur diachronique à la ressource, révélant, par exemple, quels termes archaïques ou dialectaux attestés par Ferrandi sont entrés aujourd'hui dans le somali standard, et quels sont les termes qui ont ensuite changé de sens au fil du temps.

Remerciements

Ce travail a été réalisé dans le cadre d'un accord entre le Consiglio Nazionale delle Ricerche - Istituto di Linguistica Computazionale - et la Fondazione RUT.

Références

[1] A.M. Gavello, Ugo Ferrandi, esploratore novarese, Tipografia La Cupola, Novara, 1975.
[2] E. Marini, Un novarese in Somalia: Ugo Ferrandi (1852-1928): una documentazione inedita, Bollettino storico per la provincia di Novara 82 (1991).
[3] V. Y. Mudimbe, The Invention of Africa: Gnosis, Philosophy, and the Order of Knowledge, Indiana University Press, Indianapolis, 1988.
[4] G. G. Spivak, The Rani of Sirmur: an essay in reading the archives, History and Theory 24, 3 (1985) 247-272.
[5] J. Saeed, Central Somali: A grammatical outline, Afroasiatic linguistics 8:2, Undena, Malibu, 1982.
[6] B. W. Andrzejewski, The Role of Broadcasting in the Adaptation of the Somali Language to Modern Needs. in: W. H. Whiteley (Ed.), Language Use and Social Change: Problems of Multilingualism with Special Reference to Eastern Africa. Studies Presented and Discussed at the Ninth International African Seminar at University College, Dar es Salaam, Oxford University Press, London, 1971, pp. 262-73.
[7] B. W. Andrzejewski and I. M. Lewis, Somali Poetry: An Introduction, Clarendon Press, Oxford, 1964.
[8] M. M. Moreno, Il Somalo della Somalia. Grammatica e Testi, Istituto Poligrafico Dello Stato, Roma, 1955.
[9] M. D. Wilkinson, et alii, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3(1) (2016) 1-9.
[10] Ch. Chiarcos, S. Nordhoff, and S. Hellmann, Linked Data in Linguistics, Springer, Heidelberg, 2012.
[11] J. Musse Jama, A Syntactically Annotated Corpus of Somali Literature, PhD thesis [unpublished], University of Naples "L'Orientale", 2016. See www.somalicorpus.com.

[12] A. Biswas, R. Menon, E. van der Westhuizen, Th. Niesler, Improved low-resource Somali speech recognition by semi-supervised acoustic and language model training, arXiv preprint arXiv:1907.03064, 2019.
[13] J. Laryea, N. Jayasundara, Automatic Speech Recognition System for Somali in the interest of reducing Maternal Morbidity and Mortality, Thesis, Högskolan Dalarna, Mikrodataanalys, 2020. http://urn.kb.se/resolve?urn=urn:nbn:se:du-34436.
[14] A. M. Badel, T. Zhong, W. Tai, F. Zhou, Somali Information Retrieval Corpus: Bridging the Gap between Query Translation and Dedicated Language Resources, in: H. Bouamor, J. Pino, and K. Bali (Ed.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, 7463-7469.
[15] S. Desprès, S. Szulman, Réseau terminologique versus Ontologie, in: Actes de la deuxième conférence Toth, Annecy, 5 et 6 juin 2008, pp. 17-34.
[16] R. Costa, Terminology and Specialised Lexicography: two complementary domains, Lexicographica 29 (2013) 29-42.
[17] C. Santos, R. Costa, Domain specificity, in: H. J. Kockaert, F. Steurs (Eds.), Handbook of terminology Vol. 1, John Benjamins Publishing Company, Amsterdam/ Philadelphia, 2015, pp.153-179.
[18] Ch. Roche, M. Papadopoulou, Mind the Gap: Ontology Authoring for Humanists, in: 1st International Workshop for Digital Humanities and their Social Analysis (WODHSA)-Episode V: The Styrian Autumn of Ontology, September 23-25, a Workshop hosted by Joint Ontology Workshops, Medical University of Graz (Austria), September 23-25, 2019
[19] R. Temmerman, Units of understanding in Sociocognitive Terminology studies, in: P. Faber, M-C. L'Homme (Eds.), Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge, John Benjamins, Amsterdam/Philadelphia, 2022, pp. 331–352.
[20] M. Diki-Kidiri (Ed.), Le vocabulaire scientifique dans les langues africaines, Pour une approche culturelle de la terminologie, Karthala, Paris, 2008.
[21] P. Cimiano, J. P. McCrae, and P. Buitelaar, Lexicon Model for Ontologies: Community Report. W3C Community Group Final Report, 2016. URL: https://www.w3.org/2016/05/ontolex/
[22] A. Lenci et alii, SIMPLE: A General Framework for the Development of Multilingual Lexicons, International Journal of Lexicography, XIII (4), (2000), pp. 249-263.
[23] S. Piccini, N. Ruimy. E. Giovannetti, Le lexique électronique de la terminologie de Ferdinand de Saussure : une première, in : D. Trotter, A. Bozzi, C. Fairon (Eds.), Actes du XXVIIe Congrès international de linguistique et de philologie romanes, Nancy, 15-20 juillet 2013. Section 16 : Projets en cours ; ressources et outils nouveaux. Nancy, ATILF, 255-267. http://www.atilf.fr/cilpr2013/actes/section-16.html
[24] J. Pustejovsky, The Generative Lexicon, The MIT Press, Cambridge MA, 1995