Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

TEMITALC : Text Mining et TAL pour Analyser le Langage des Cachalots

José Coch[1] Olivier Adam[2]
(1) Service NLP, Dassault Systèmes, 10, place de la Madeleine, 75008 Paris, France
(2) Institut d'Alembert Sorbonne Université, 4 place Jussieu 75005 Paris France
jose.cochdiyacovo@3ds.com, olivier.adam@sorbonne-universite.fr

RESUME

Les cachalots sont des cétacés qui communiquent par des clics organisés en séquences appelées "codas". Elles sont numérisables relativement facilement : il existe en effet des corpus de transcriptions de conversations.

Une collaboration interdisciplinaire entre le Service NLP de Dassault Systèmes et l'équipe Bioacoustique de Sorbonne Université, a initié un projet d'application des techniques du TAL au cas des cachalots. Ses premiers résultats sont exposés dans les Actes de TextMine'23.

Le projet utilise le logiciel Proxem, qui permet de construire des modèles de langue à partir des corpus à analyser.

TEMITALC couvre les points suivants :

- Analyse des propriétés formelles du langage,
- Identification des corrélations entre des éléments non linguistiques et des éléments du langage.

Il bénéficie d'un financement de Dassault Systèmes et de Sorbonne Université. Sa fin est prévue pour décembre 2024.

Nos résultats vont contribuer à décrire le sophistiqué langage d'une espèce non-humaine.

ABSTRACT

TEMITALC: Text Mining and NLP to Analyze the Language of Sperm Whales

Sperm whales are cetaceans that communicate through clicks organized in sequences called "codas". They can be digitized relatively easily: there are indeed corpora of transcriptions of conversations.

An interdisciplinary collaboration between the NLP Service of Dassault Systèmes and the Bioacoustics team of Sorbonne University, initiated a project to apply NLP techniques to the case of sperm whales. Its first results are set out in the Proceedings of TextMine'23.

The project uses the Proxem software, which makes it possible to build language models from the corpora to be analyzed.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 6 : projets, pages 23–25.

TEMITALC covers the following points:

- Analysis of the formal properties of this language,
- Identification of correlations between non-linguistic elements and language elements.

It receives funding from Dassault Systèmes and Sorbonne University. Its end is scheduled for December 2024.

Our results will help to describe the sophisticated language of a non-human species.

MOTS-CLES : zoolinguistique ; text-mining ; interdisciplinaire ; ordre des mots ; corrélations sémantiques.
KEYWORDS: zoolinguistics; text-mining; interdisciplinary; word order; semantic correlations

1. TEMITALC : Text Mining et TAL pour Analyser le Langage des Cachalots

Les cachalots (Physeter macrocephalus) sont les plus grands des cétacés à dents. Comme tous les cétacés, ils communiquent notamment par des émissions vocales. Les cachalots produisent des clics au cours de leurs activités vitales et leurs interactions sociales. Certains de ces sons sont organisés en séquences temporelles, appelées « codas ». Depuis plus d'une dizaine d'années, des échanges audio ou « conversations » entre cachalots sont enregistrés dans de nombreux endroits dans le monde, par exemple dans l'Océan Pacifique, dans les Caraïbes et dans l'Océan Indien. La particularité des échanges vocaux entre cachalots fait que ces codas sont numérisables relativement facilement. Ainsi, il existe des corpus de transcriptions de conversations en particulier venant des origines géographiques citées.

Durant 2022, une collaboration entre le Service NLP de Dassault Systèmes et l'équipe Bioacoustique de Sorbonne Université, basée sur les enregistrements sonores collectés et mis à disposition par Longitude 181 et Label Bleu Production, nous a permis d'initier un projet d'application des techniques de Text Mining et Traitement Automatique du Langage à l'étude du langage des cachalots. Nous avons exposé les premiers résultats du projet dans un article publié dans les Actes de l'atelier TextMine'23 de la conférence EGC'2023 concernant un corpus de cachalots résidents au large de l'Ile Maurice et identifiés individuellement.

Nous utilisons dans ce projet le logiciel Proxem Studio, qui a la particularité de pouvoir être appliqué sans modèle de langue préalable car il peut construire des modèles de langue à partir des corpus à analyser.

L'objectif du projet couvre les points suivants :

- Optimiser et automatiser la transcription en codas des échanges audio entre cachalots,
- Analyser les propriétés formelles du langage des cachalots : mettre en évidence que l'ordre entre codas a une importance, et découvrir s'il est possible de décrire une proto-syntaxe de ce langage,

- Mettre au point un référentiel d'éléments non linguistiques (comportements sociaux, données démographiques, relations familiales) et identifier des codas ou des séquences de codas montrant une corrélation avec ces éléments non linguistiques, et in fine, avancer des hypothèses sur la fonction de certaines codas ou séquences de codas,
- Etudier les corrélations entre les participants à chaque conversation et les codas émis afin de déterminer si des codas ou séquences de codas peuvent être associées à des individus.

Le projet bénéficie d'un financement de Dassault Systèmes et de Sorbonne Université. La fin du projet est prévue pour décembre 2024.

Nos résultats vont contribuer ainsi à décrire le sophistiqué langage d'une espèce non-humaine.

Références

ADAM, O., A. YERNAUX, M. SAUVÊTRE, J. NGOSSO, G. NUEL, M. HAFFNER-TRINH, R. TROUSSIER, Z.-L. GUILLERM, L. PICON, L. BARLUET, J. MACKY, L. BARLUET DE BEAUCHESNE, V. KUHN, F. DELFOUR, V. SARANO, H. VITRY, A. PREUD'HOMME, R. HEUZEY, J.-L. JUNG, ET F. SARANO (2020). Study of behaviours and emitted codas during sperm whale social interactions. *e-Forum Acusticum 2020*, Dec 2020, Lyon, France. pp.3225-3227.

ANDREAS, J., G. BEGUS, M. BRONSTEIN, R. DIAMANT, D. DELANEY, S. GERO, S. GOLDWASSER, D. GRUBER, S. HAAS, P. MALKIN, R. PAYNE, G. PETRI, D. RUS, P. SHARMA, P. TØNNESEN, A. TORRALBA, D. VOGT, ET R. WOOD (2021). Cetacean translation initiative: a roadmap to deciphering the communication of sperm whales. arXiv (2104.08614)

BERMANT, P., M. BRONSTEIN, R. WOOD, S. GERO, ET D. GRUBER (2019). Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. *Scientific Reports* 9, 1–10.

BOSSHARD, A., M. LEROUX, N. LESTER, B. BICKEL, S. STOLL, ET S. TOWNSEND (2022). From collocations to call-ocations: using linguistic methods to quantify animal call combinations. BEHAVIORAL ECOLOGY AND SOCIOBIOLOGY ; Heidelberg.

COCH, J., BERKENBAUM, L., ADAM, O. (2023). Une contribution du Text-mining à la connaissance du langage des cachalots. In *Actes de TextMine, atelier de la conférence EGC (Extraction et Gestion des Connaissances)*, pp. 15-32. Lyon, France, janvier 2023.
https://textmine.sciencesconf.org/data/pages/TextMine23.pdf

DOH, YANN & ECALLE, BEVERLEY & DELFOUR, FABIENNE & PANKOWSKI, CYPRIEN & COZANET, GILDAS & BECOUARN, GUILLAUME & OVIZE, MARION & DENIS, BERTRAND & ADAM, O.. (2023). Performance Assessment of the Innovative Autonomous Tool CETOSCOPE© Used in the Detection and Localization of Moving Underwater Sound Sources. *Journal of Marine Science and Engineering*. 11. 960. 10.3390/jmse11050960.