Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Comment j'ai recontré ton *snowclone* : Découverte non-supervisée de moules de *snowclones* dans de grands jeux de données

Julien Bezançon[1, 2] Gaël Lejeune[1] Marceau Hernandez[1]
(1) STIH, CERES, Sorbonne Université, 75006, Paris, France
(2) LISN, CNRS, Université Paris-Saclay, 91400, Orsay, France
(1)bezancon@lisn.fr
(2)prenom.nom@sorbonne-universite.fr

RÉSUMÉ

Un *snowclone* est un moule d'expression multi-mots qui inclut des positions flexibles, acceptant des variations. Une caractéristique propre aux *snowclones* est que l'expression ayant servie à leur création reste identifiable et que son sens est conservée dans le nouvel énoncé créé. Cependant, il n'a jamais été démontré que les substitutions sont limitées aux positions flexibles prédéfinies au sein d'un *snowclone*. En utilisant le *locality sensitive hashing*, nous extrayons automatiquement des *snowclones* depuis le jeu de données non-commercial d'IMDb pour créer FROST, le premier lexique multilingue de *snowclones*. Ce lexique comprend 30 826 *snowclones* et 1 059 824 énoncés créés à partir de ces *snowclones* répartis dans 30 langues. Nous avons aussi annoté 1 500 *snowclones* et 1 000 énoncés pour évaluer la qualité du lexique créé. Ce travail révèle que la plupart des substitutions au sein d'un *snowclone* s'opèrent à des positions prédéfinies.

ABSTRACT

How I Met Your Snowclone : Unsupervised Discovery of Snowclone Patterns in Large Datasets

Snowclones are a type of Multiword Expression that includes open slots, i.e. positions that can be filled with various words. A key feature of snowclones is that the original MWE remains recognizable, conveying its meaning into the new form. However, previous work has not shown whether such substitutions are limited to fixed positions. We propose to use Locality Sensitive Hashing to automatically extract snowclone patterns from the non-commercial IMDb dataset. We create the FROST lexicon, comprising 30,826 pattern candidates and 1,059,824 snowclone candidates distributed in 30 languages. We annotated 1,500 discovered patterns and 1,000 snowclones to assess its quality. Our findings suggest that most substitutions in snowclones occur at consistent positions.

MOTS-CLÉS : snowclone, expressions multi-mots, locality sensitive hashing, lexique.

KEYWORDS: snowclone, multiword expressions, locality sensitive hashing, dataset, lexicon.

ARTICLE ACCEPTÉ À : Language Resources and Evaluation Conference (LREC), 2026.
URL : https://lrec.elra.info/lrec2026-main-622

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 5–5.