UniDive
2nd General Meeting
University of Naples "L'Orientale
Naples, Italy, 8-9 February 2024
https://unidive.lisn.upsaclay.fr/

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

Annotating French MWEs for French L2 learning

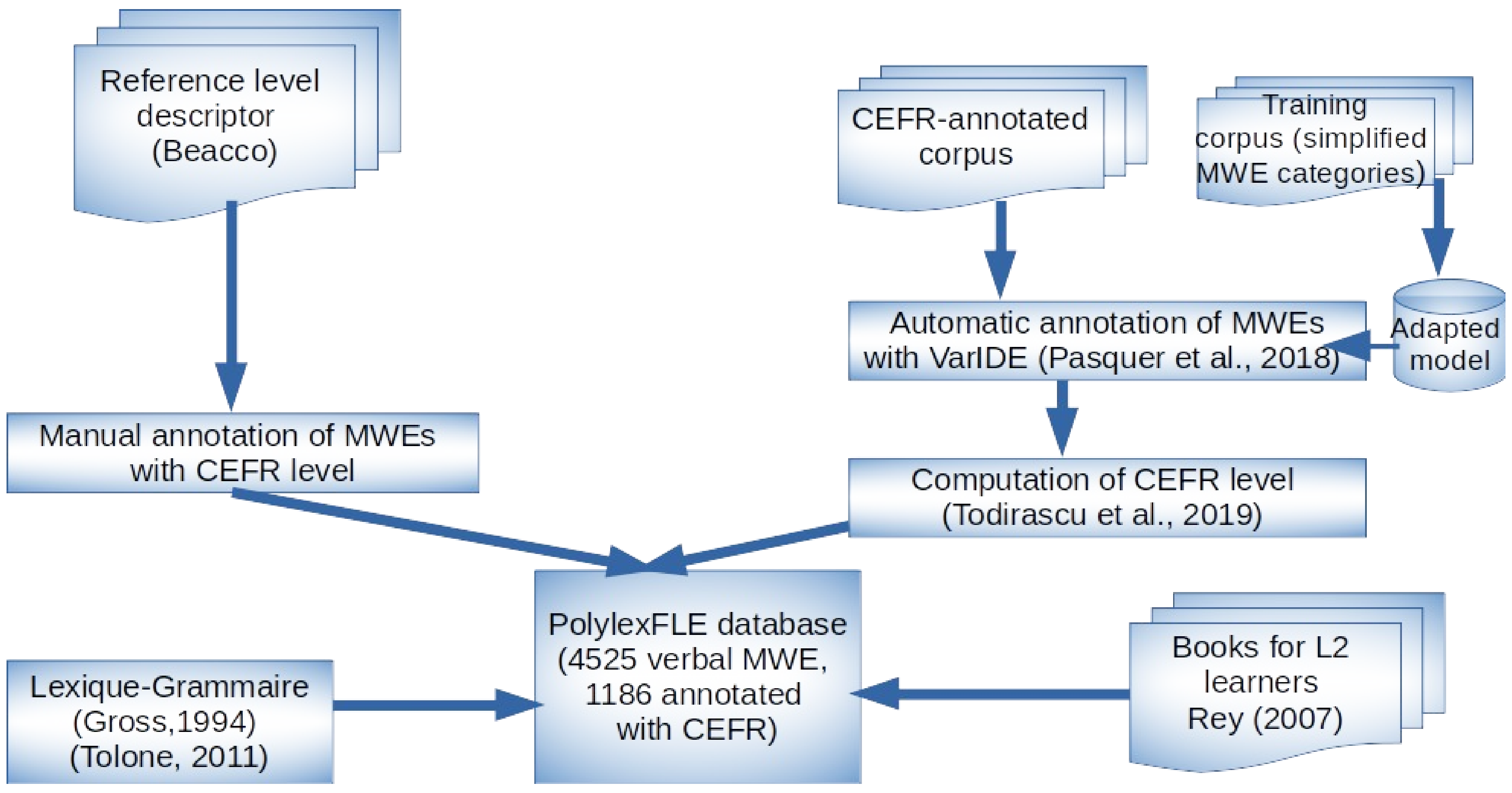
Amalia Todirascu
LiLPa, University of Strasbourg
WG1 (WG1.2)

Goals

1) create a CEFR-level corpus, containing verbal MWEs, automatically annotated with VarIDE (Pasquer et al., 2018)
2) create a lexical database of MWEs annotated with CEFR level

Motivation and context

- Knowledge of MWE is crucial for L2 learners (Bahn and Eldaw, 1993)
- language acquisition depends on the category of MWEs (Siyanova, 2017)
- Few French digital ressources annotated with CEFR level : (Alfter and Graën, 2019) or FLELex (Tack *et al.*, 2016), but few MWEs available
- French Reference level descriptor : paper format (Beacco and Porquier, 2007), (Beacco, 2008), (Beacco and Porquier, 2008), (Beacco *et al.*, 2011), (Beacco *et al.*, 2004) : few MWEs described
- Context: the ANR project STAR-FLE (2024-2028)


Reference level descriptor (Beacco)
CEFR-annotated corpus
Training corpus (simplified MWE categories)
Automatic annotation of MWEs with VarIDE (Pasquer et al., 2018)
Adapted model
Manual annotation of MWEs with CEFR level
Computation of CEFR level (Todirascu et al., 2019)
PolylexFLE database (4525 verbal MWE, 1186 annotated with CEFR)
Books for L2 learners Rey (2007)
Lexique-Grammaire (Gross,1994) (Tolone, 2011)

Adapting VarIDE

- VarIDE (Pasquer et al, 2018) identifies MWEs and variants
- Initially provides PARSEME categories
- Annotating the training and test corpus with our categories according to our guidelines
- Mapping the categories
  - VID => idioms, if the semantic criterion is satisfied
  - LVC.full => either collocations (Mel'cuk, 1998) or fixed expressions (Gross, 1993)

Definition and categories

- MWEs are sequences of words, which might be discontinous, presenting at least two lexical, statistical, syntax or semantic idiosyncracy (Constant *et al.*, 2017)
- **Specific types of verbal MWEs (difficult for L2 learners)** : idioms, collocations, fixed expressions

Idioms :
-non-compositionality: *mettre les pieds dans les plats* 'to put your feet in it';
-fixed/no determiner,
-impossible syntactic transformations

Collocations :
-strong lexical preferences (*poser une question*, but not **demander une question* 'ask a question');
-the sense is more compositional
-syntactic variation

Fixed expression:
expressions with a head verb (*être sans reproche* 'to be without reproach',
être d'accord 'to agree'), but the object is fixed and lexicalized

Corpus for MWE extraction

- 324.545 words, distributed among 6 CEFR levels: A1 (15.620 words), A2(43.422 words), B1(57.795 words), B2(101.361 words), C1(54.057 words) and C2(52.290 words)

Guidelines

- Guidelines explaining lexical, syntactic, semantic criteria to distinguish the three categories
- Test the guidelines 30 representative texts of the A1 to C2 levels
- 3 annotators and intercoder agreement : 272 MWEs are manually annotated

Agreement	Measure	Average	Coder1 – coder 2	Coder 2 – Coder 3	Coder 3 – Coder 1
Delimitation	F-measure	0.75	0.756	0.774	0.733
Delimitation	κ	0.74	0.74	0.76	0.71
Categories	κ	0.63	0.70	0.61	0.59

Conclusion and further work

- Annotation corpus for VarIDE in progress
- Comparing criteria with PARSEME guidelines
- Guidelines applied on a small part of the corpus
- intercoder agreement to be improved, for categories

References
Bahns J. & Eldaw M. (1993). Should We Teach EFL Students Collocations? System, 21(1), p. 101-14.
Beacco, J.-C. & Porquier, R. (2008). Niveau A2 pour le français : utilisateur-apprenant élémentaire, Didier, Paris.
Beacco, J.-C., Bouquet, S., Porquier, R. (2004). Niveau B2 pour le français : un référentiel : utilisateur-apprenant indépendant, Didier, Paris.
Beacco, J.-C. (2008). Niveau A1/A2 pour le français: Textes et références. Didier.
Constant, M., Ergÿgit G., Monti, J., Van der Plas, L., Ramisch C., Rosner M., Todirascu A. (2017). Multiword Expression Processing : A Survey. Computational Linguistics, 43(4), p. 837–892.
François T., Gala N., Watrin P. & Fairon C. (2014). FLELex : a graded lexical resource for French foreign learners. In Proc. of the Language and Resources Evaluation Conference (LREC 2014), Reykjavick, Iceland, p. 3766–3773.
Gross, M. (1994). Constructing Lexicon-Grammars, In Atkins, R. and Zampolli, A., Computational approaches to the lexicon, Oxford Univ. Press, p. 213-263
Mel'čuk, I. (1998). Collocations and lexical functions. In Phraseology. Theory, analysis, and applications (p. 23–53). Citeseer.
Rey, I. G. (2007). La didactique du français idiomatique. Editions Modulaires Européennes InterCommunication.
Siyanova-Chanturia, A. (2017). Researching the teaching and learning of multi-word expressions. Language Teaching Research, 21(3), 289□297.
Tack, A., François, T., Ligozat, A.-L., & Fairon, C. (2016). Evaluating lexical simplification and vocabulary knowledge for learners of French: possibilities of using the FLELex resource. Proceedings of LREC 2016), 230□236.
Todirascu, A., Cargill, M., Francois, T. (2019). PolylexFLE : une base de données d'expressions polylexicales pour le FLE. Actes de la 26e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), Toulouse, France, p. 143-156.
Tolone, E. (2011). Maintenance du Lexique-Grammaire : Formules définitoires et arbre de classement. Ressources Linguistiques Libres, 52(3), 153□190.

lilpa linguistique, langues, parole
Université de Strasbourg
University of Strasbourg, 14 rue René Descartes, 67084 Cedex, France
todiras@unistra.fr

anr