International Corpus of Learner English

Version 2

Sylviane Granger, Estelle Dagneaux, Fanny Meunier,
Magali Paquot (eds.)

UCL PRESSES
UNIVERSITAIRES
DE LOUVAIN

Registration of copyright: D/2009/9964/3

ISBN : 978-2-87463-143-6
Cover : mikedsign
Printed in Belgium

Distribution : www.i6doc.com, online university publishers

Available on order from bookshops or at
CIACO University Distributors
Grand-Place, 7
1348 Louvain-la-Neuve, Belgium
Tel. 32 10 47 33 78
Fax 32 10 45 73 50
duc@ciaco.com

PREFACE TO FIRST EDITION

Although it has not been around for long, the learner corpus field has already opened up exciting avenues and yielded some interesting results. However, major developments in the field are seriously hindered by the lack of availability of learner corpora. It is our hope that the release of the *International Corpus of Learner English* will mark a new stage in the evolution of EFL research. The corpus, which is the result of over 10 years of international collaboration between a large number of universities, contains 2.5 million words of English written by learners from 11 different mother tongue backgrounds. It is now being released in CD-ROM format with a user-friendly interface which allows researchers to compile their own tailor-made corpora on the basis of a set of predefined criteria. It is accompanied by a handbook which contains a comprehensive description of the corpus, a detailed user manual and brief articles describing the status of English in the countries of origin of the learners. We hope that it will be useful both to SLA theoreticians and EFL practitioners. We are convinced that it can contribute to giving theories of second language acquisition a more solid empirical foundation and lead to the production of more learner-aware pedagogical material, designed for advanced EFL learners in general or focused on the needs of one specific national learner population.

Estelle Dagneaux
Sylviane Granger
Fanny Meunier

December 2002

PREFACE TO SECOND EDITION

The first edition of the *International Corpus of Learner English (ICLE)*, published in 2002, has been very well received internationally. It has been used in a wide range of research projects, served as the basis for many MA and PhD theses and generally played a key role in promoting the emerging field of learner corpus research. The second version differs from the first not only in terms of the higher amount and greater diversity of the learner data included, but also by the improved functionalities it affords. The current version is much more than a simple database of texts: it contains a built-in concordancer which allows for both simple and more complex searches. Another major change is that the rich learner profile information attached to each text is now linked to the search results. In developing the new version, we have taken into account the abundant feedback we have received over the years from scholars using the corpus. We hope the new version will go some way to meeting their requirements. The *ICLE* is a dynamic project: the corpus is regularly updated as new varieties of learner English are added, so there is little doubt that there will be a third release. We therefore encourage users to contact us with their feedback and suggest additional features for future editions.

Estelle Dagneaux
Sylviane Granger
Fanny Meunier
Magali Paquot

February 2009

ACKNOWLEDGEMENTS

First and foremost, we wish to express our deep gratitude to the *ICLE* national coordinators for their relentless efforts in collecting data and the wonderful collaboration we have enjoyed over so many years. A long-term international project such as *ICLE* does not only result in scientific output. It also creates bonds of friendship which will outlast the life of the project. Our thanks also go to the members of the *Centre for English Corpus Linguistics (CECL),* not only for their precious collaboration at various stages of the project but also for their unfailing enthusiasm. We have also benefited from the dedication and conscientious efforts of numerous students who have contributed to the many painstaking yet necessary tasks that corpus collection entails. Special thanks are due to the beta-testers of *ICLEv2* – Gaëtanelle Gilquin, Przemek Kaszubski and Signe Oksefjell Ebeling – who have provided many useful suggestions for improvement of the interface. We also wish to express our profound gratitude to Paul Rayson, who kindly agreed to POS-tag all the *ICLE* files with the CLAWS7 tagger. Last but not least, we are deeply indebted to the many – well over four thousand – learners who have agreed to contribute their essays to the corpus.

The *ICLE* project would never have got off the ground without the initial and continuing encouragement from three major figureheads in the corpus linguistics field: Jan Aarts, the late Sidney Greenbaum and Geoffrey Leech. By taking a keen interest in the project at a time when the very notion of learner corpus did not even exist, they gave us the necessary confidence to move forward. We gratefully dedicate this volume to them.

We are also greatly indebted to the different funding bodies which have supported the project over the years. The University of Louvain Research Fund played a crucial role in providing us with the initial resources to launch the project. We have also benefited from the generous support of the Belgian National Scientific Research Fund at various stages of the project and from a special research grant from the Walloon Region.

Last but not least, we wish to express warm thanks to Cédrick Fairon, Director of the *Centre de Traitement Automatique du Langage*

(CENTAL) of the University of Louvain, who welcomed the idea of joining forces on the production of the CD-ROM and helped us to bring the project to a successful conclusion. Special thanks are due to Claude Devis from the *CENTAL* for the wonderful job he did in setting up the *ICLE* database and developing a user-friendly interface for the CD-ROM as well as for his unfailing kindness and patience.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

LIST OF ABBREVIATIONS

CA	Contrastive Analysis
CIA	Contrastive Interlanguage Analysis
CLAWS	Constituent Likelihood Automatic Word-tagging System
EA	Error Analysis
EFL	English as a Foreign Language
ELT	English Language Teaching
ESL	English as a Second Language
FL	Foreign language
ICLEv1	First version of the ICLE CD-ROM and Handbook (Granger, Dagneaux & Meunier eds. 2002)
ICLEv2	Second version of the ICLE CD-ROM and Handbook
L1	Native language
L2	Foreign/Second language
LOCNESS	Louvain Corpus of Native English Essays
POS	Part-of-speech
SLA	Second Language Acquisition
WST	WordSmith Tools (text retrieval program)

List of key terms provided in the statistics windows of *ICLEv2*

Corpus	the whole *ICLEv2* corpus. It consists of 6,085 texts and totals 3,753,030 words.
Sub-corpus	the number of profiles and corresponding texts matching the user's corpus compilation criteria before any potential deselection
Selected Corpus	the number of profiles and corresponding texts left after deselection of profiles (if any)
Result Selected Corpus	the number of texts in the Selected Corpus which contain a linguistic query entered by the user
CorpVar	all texts in the Selected Corpus that match one specific variable

0. INTRODUCTION

This combined CD-ROM and handbook represents the culmination of a project that started in 1990. At the time, corpus linguistics was already a well-established linguistic methodology which was showing its full potential in the field of variation studies. Using a combined quantitative/qualitative approach, corpus linguists were providing much more accurate descriptions of varieties of English than had ever been available before. Although most native varieties of English - regional, diachronic, stylistic - benefited from this new corpus approach, the non-native varieties were completely neglected, which seems strange and unjustifiable when one considers the fact that the number of non-native speakers of English in the world far exceeds that of native speakers. The project launched by Sylviane Granger at the Université catholique de Louvain in October 1990 aimed to bridge that gap. Initially, the project focused exclusively on the writing of advanced French-speaking learners of English, but the idea quickly caught on. Other EFL varieties were added and the project became known as the *International Corpus of Learner English (ICLE).* A first CD-ROM was released in 2002 (Granger, Dagneaux & Meunier eds.). It contained written data produced by learners from 11 different mother tongue backgrounds: Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish and Swedish. From the start we were keen to follow Sinclair's (1991) guidance that "a corpus should be as large as possible, and should keep on growing" and data collection continued uninterruptedly after the release of the first CD-ROM – indeed is still continuing today, with several new national subcorpora still under construction.

With 4.5 million words, the current version of the corpus contains over 1 million words more than the first. The geographical spread of the project has also widened: while the first version only contained data from Europe, the second version includes data from China, Japan and South Africa, in addition to two new European subcorpora (Norway and Turkey). Users of *ICLEv2* will thus benefit from a larger empirical basis for their research. More importantly, they will be able to carry out their research in a totally new technological environment. In *ICLEv1*, users used the CD-ROM to select learner texts that matched a set of predefined attributes but had to use other tools to query the data. In *ICLEv2* the linguistic analysis can be carried out without having to leave the CD-ROM's environment. The built-in

concordancer allows for both simple lexical searches and more sophisticated searches using a POS-tagged version of the data. A major advantage of this integration is that the learner profile information, which was detached from the compiled learner corpora in *ICLEv1*, remains available at all times in the second version, allowing for interesting links between search results and learner/task variables.

The body of this handbook is subdivided into six sections. Part I gives a general overview of the corpus, i.e. the design criteria, the task and learner variables, markup and linguistic annotation. Parts II and III introduce the 16 national projects: the teams that have collected the data (Part II) and the national subcorpora (Part III). Part IV provides some guidelines on how to use the corpus to carry out research into learner language. Part V contains a detailed user manual. Part VI provides information on the status of English in the learners' country of origin.

To order: http://www.i6doc.com/en/collections/cdicle/

- Single user licence
- Multiple-user licence (2-10)
- Multiple-user licence (11-25)