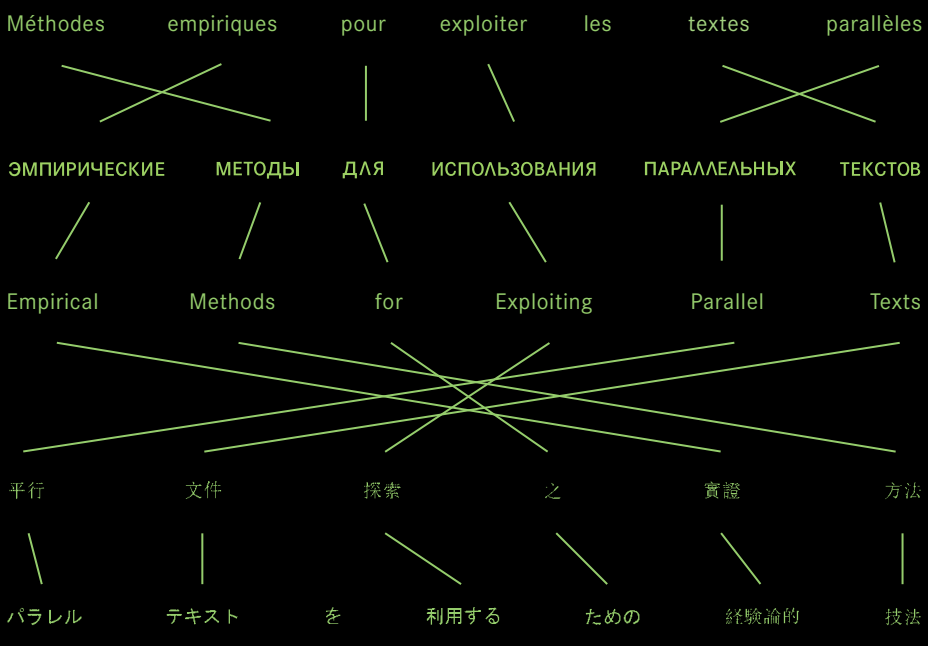
I. DAN MELAMED

EMPIRICAL METHODS FOR EXPLOITING PARALLEL TEXTS



Empirical Methods for Exploiting Parallel Texts

Empirical Methods for Exploiting Parallel Texts

I. Dan Melamed

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Times Roman by Windfall Software using ZzTEX and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Melamed, I. Dan.
Empirical methods for exploiting parallel texts / I. Dan Melamed
p. cm.
A thorough revision of the author's thesis (Ph.D.—1998).
Includes bibliographical references and index.
ISBN 0-262-13380-6 (hc.: alk. paper)
1. Machine translation. 2. Linguistic models. I. Title.
P309.M45 2001
418′.02—dc21 00–038028

for Flora and Assir Melamed—my rock and my foundation

Contents

	Acknowledgments	xi
1	**Introduction**	1
I	**TRANSLATIONAL EQUIVALENCE AMONG WORD TOKENS**	5
2	**A Geometric Approach to Mapping Bitext Correspondence**	7
2.1	Introduction	7
2.2	Bitext Geometry	8
2.3	Previous Work	9
2.4	The Smooth Injective Map Recognizer (SIMR)	13
	2.4.1 Overview	13
	2.4.2 Point Generation	14
	2.4.3 Noise Filter	17
	2.4.4 Point Selection	18
	2.4.5 Reduction of the Search Space	19
	2.4.6 Enhancements	21
2.5	Parameter Optimization	23
2.6	Evaluation	24
2.7	Implementation of SIMR for New Language Pairs	30
	2.7.1 Step 1: Construct Matching Predicate	30
	2.7.2 Step 2: Construct Axis Generators	31
	2.7.3 Step 3: Reoptimize Parameters	32
2.8	Conclusion	33
3	**Application: Alignment**	35
3.1	Introduction	35
3.2	Correspondence is Richer Than Alignment	35
3.3	The Geometric Segment Alignment (GSA) Algorithm	37
3.4	Evaluation	38
3.5	Conclusion	40
4	**Application: Automatic Detection of Omissions in Translations**	41
4.1	Introduction	41
4.2	The Basic Method	41
4.3	Noise-Free Bitext Maps	43
4.4	A Translator's Tool	44

4.5 Noisy Bitext Maps 45
4.6 ADOMIT 46
4.7 Simulation of Omissions 48
4.8 Evaluation 49
4.9 Conclusion 53

II THE TYPE-TOKEN INTERFACE 55

5 Models of Co-occurrence 57
5.1 Introduction 57
5.2 Relevant Regions of the Bitext Space 58
5.3 Co-occurrence Counting Methods 59
5.4 Language-Specific Filters 62
5.5 Conclusion 63

6 Manual Annotation of Translational Equivalence 65
6.1 Introduction 65
6.2 The Gold-Standard Bitext 66
6.3 The Blinker Annotation Tool 68
6.4 Methods for Increasing Reliability 69
6.5 Inter-Annotator Agreement 72
6.6 Conclusion 77

III TRANSLATIONAL EQUIVALENCE AMONG WORD TYPES 79

7 Word-to-Word Models of Translational Equivalence 81
7.1 Introduction 81
7.2 Translation Model Decomposition 82
7.3 The One-to-One Assumption 86
7.4 Previous Work 87
7.4.1 Non-Probabilistic Translation Lexicons 87
7.4.2 Re-estimated Sequence-to-Sequence Translation Models 89
7.4.3 Re-estimated Bag-to-Bag Translation Models 93
7.5 Parameter Estimation 94
7.5.1 Method A: The Competitive Linking Algorithm 97

7.5.2 Method B: Improved Estimation Using an Explicit Noise Model 99
7.5.3 Method C: Improved Estimation Using Pre-Existing Word Classes 103
7.6 Effects of Sparse Data 104
7.7 Evaluation 107
7.7.1 Evaluation at the Token Level 107
7.7.2 Evaluation at the Type Level 114
7.8 Application to MT Lexicon Development 119
7.9 Conclusion 121

8 Automatic Discovery of Non-Compositional Compounds 123
8.1 Introduction 123
8.2 Objective Functions 124
8.3 Search 126
8.4 Predictive Value Functions 127
8.5 Iteration 128
8.6 Credit Estimation 131
8.7 Single-Best Translation 134
8.8 Experiments 135
8.9 Related Work 143
8.10 Conclusion 144

9 Sense-to-Sense Models of Translational Equivalence 147
9.1 Introduction 147
9.2 Previous Work 148
9.3 Formulation of the Problem 150
9.4 Noise Filters 151
9.5 The SenseClusters Algorithm 152
9.6 An Application 154
9.7 Experiments 155
9.7.1 Quantitative Results 156
9.7.2 Qualitative Results 160
9.8 Conclusion 162

10 Summary and Outlook 165

A Annotation Style Guide for the Blinker Project 169
A.1 General Guidelines 169
A.1.1 Omissions in Translation 170
A.1.2 Phrasal Correspondence 171
A.2 Detailed Guidelines 173
A.2.1 Idioms and Near Idioms 173
A.2.2 Referring Expressions 175
A.2.3 Verbs 177
A.2.4 Prepositions 178
A.2.5 Determiners 180
A.2.6 Punctuation 180

Notes 183
References 187
Index 193

Acknowledgments

This book is a thorough revision of my 1998 Ph.D. dissertation. As with all dissertations, this one owed much to people who went out of their way to help me learn. I am particularly grateful to:

- Mitch Marcus, for creating the best imaginable research environment and for teaching me how to use it.
- My dissertation committee, for preventing me from making a fool of myself in print.
- The LINC lab empiricist gang. If I count carefully, I might find that half of the ideas in this book were germinated by Jason Eisner, and that the rest were proposed by Mike Collins, Adwait Ratnaparkhi, Lyle Ungar and David Yarowsky.
- Innumerable past and present students, staff, post-docs, visitors and faculty at the Institute for Research in Cognitive Science at the University of Pennsylvania and its member departments, who came to my CLiFF talks, read drafts of my papers, and helped me refine my crazy ideas into ideas that were not merely crazy but also useful.
- My collaborators and advisors outside of Penn. I have been very lucky to work with Pierre Isabelle and his team (now at University of Montreal), with Philip Resnik at Sun Labs and at UMIACS, and with Young-Suk Lee of MIT Lincoln Labs. Some of the ideas in this book can be directly traced to correspondence with Ken Church at AT&T Research, Ido Dagan at Bar-Ilan University, George Foster at University of Toronto, and Djoerd Hiemstra at University of Twente.
- My undergraduate mentor Graeme Hirst, for encouraging my nascent interest in computational linguistics, thereby propelling me towards an exciting research career.
- Peter Jackson and my colleagues in the Computer Science Research Department of West Group, for their support and encouragement during the rewriting and editing stages.
- The anonymous reviewers who read the manuscript so carefully and who gave me many new insights.

Empirical Methods for Exploiting Parallel Texts

1 Introduction

One of the most exciting promises of research in artificial intelligence is computers that can understand natural human language. The main obstacle to fulfilling this promise has been the difficulty of modeling linguistic phenomena in sufficient detail. Natural language follows few hard and fast rules. Therefore, a good model must account for tendencies and likelihoods. Although people use language all the time, they cannot accurately assign probability distributions over linguistic data by introspection.

Fortunately, the amount of computing power available for research has been steadily doubling since the 1980s and there has been a dramatic increase in the amount of linguistic data available online. These resources have made possible a new approach to the language modeling problem—the empirical approach. If people cannot specify statistical language models by introspection, perhaps computers can induce the models from data.

One kind of raw material that has become much more plentiful since the birth of the Web is parallel texts in multiple languages (Resnik, 1999). A text and its translation constitute a *bitext*. Bitexts are one of the richest sources of linguistic knowledge because the translation of a text into another language can be viewed as a detailed annotation of what that text means. One might think that if that other language is also a natural language, then a computer is no further ahead, because it cannot understand the annotation any more than it can understand the original text. However, just the knowledge that the two data streams are semantically equivalent leads to a kind of understanding that enables computers to perform an important class of "intelligent" functions.

In particular, many functions that involve two or more languages can now be automated to some degree. Easier tasks, such as finding corresponding regions of parallel texts, can now be fully automated. On the other extreme of the difficulty continuum, the Web has spawned a number of online services that perform fully automatic translation from one language to another, with varying degrees of success. The Web has also created a demand for new kinds of multilingual functions, such as cross-language information retrieval, that computers are far better suited to perform than people. All these functions require knowledge of semantic equivalence across languages.

Formally, semantic equivalence between different languages or parts thereof is a mathematical relation called *translational equivalence*. The relation holds between expressions with the same meaning. The expressions can be as small as individual morphemes or as large as entire texts and speeches. To achieve the kind of limited understanding described above, it is first necessary to break down translational equivalence between texts into equivalence between smaller

text units. The present work is about automatic discovery and exploitation of translational equivalence between words.

Translational equivalence can range over token pairs or type pairs. *Tokens* are instances of linguistic units in particular positions in particular texts. *Types* are abstract sets of tokens with identical appearance. For example, the first word in this sentence is a For token. That token and all other For tokens in this book and in other English texts constitute the word type **For**. With any kind of data, a computer must know something about the properties of types to infer properties of tokens, and vice versa. Using knowledge about types to find their tokens is usually called pattern recognition. Using information about tokens to induce models of their types is called learning. This book is organized around the type-token symbiosis.

Part I of the book deals with pattern recognition—using knowledge about types to infer knowledge about tokens. Even a very rough approximation of translational equivalence among word types is sufficient to recognize translationally equivalent tokens. Chapter 2 shows how to find corresponding word tokens in bitext automatically using, e.g., only the cognate heuristic and/or a few hundred entries automatically extracted from an on-line bilingual dictionary. Chapter 3 shows how correspondence among word tokens can be quickly and accurately extended to correspondence among longer bitext segments such as sentences. Chapter 4 shows how the techniques developed in chapter 2 can be applied to build a translators' tool for automatically detecting omissions in translations. Omission detection is typical of the kind of problem whose solution was previously thought to require full understanding of two languages, but that now can be approached with bitext-driven empirical methods.

Part II of the book deals with issues at the type-token interface. Chapter 5 describes how a *model of co-occurrence* can abstract a translational equivalence relation at the token level to a translational equivalence relation at the type level. Almost all published methods of learning translational equivalence among word types, including the methods in this book, start by considering what pairs of word tokens co-occur in corresponding regions of the training bitext. Counting co-occurrences correctly is crucial, but the most commonly used counting method turns out to be suboptimal for most applications. Chapter 5 exposes the problem and offers some solutions. It also shows how to count co-occurrences in *arbitrary* bitexts, not just in the restricted class of bitexts most often addressed in the literature to date.

The other chapter in part II describes a project undertaken to manually annotate translational equivalence at the token level in a significant subset of a

large bitext. I use these annotations in subsequent chapters as a gold standard for evaluating various models of translational equivalence among word types. The annotation style guide appears as appendix A. The annotations themselves are freely downloadable for research purposes.

Part III of the book is about *models of translational equivalence* among word types (or *translation models*,[1] for short). Chapter 7 describes how to exploit two properties of bitext to improve translation model accuracy. The chapter also shows how a statistical model can incorporate various kinds of pre-existing knowledge that might be available about particular language pairs. Even the simplest kinds of language-specific knowledge, such as the distinction between content words and function words, are shown to reliably boost translation model accuracy. Chapter 8 tackles another long-standing problem: how to estimate translational equivalence, given that many word sequences are translated non-compositionally. The solution lies in an information-theoretic method for automatically discovering these non-compositional compounds and then treating them as atomic words within the methods of chapter 7. Chapter 9 develops a new method for unsupervised word-sense discrimination, in order to enable word-to-word translation models to account for polysemy.

The main innovations in this book have been rigorously evaluated and shown to advance the state of the art on the relevant criteria. Significant quantitative improvements in engineering methods often translate into qualitative improvements. Occasionally, a quantitative improvement will make a new application feasible by tipping the cost-efficiency balance, whether financial cost or computational cost. I hope that each reader will envision at least one new application of the ideas in this book, in addition to the ones I have proposed here.

Throughout the book, $\mathcal{CALLIGRAPHIC}$ letters denote text corpora and other sets of sets; CAPITAL letters denote collections, including sequences and bags; *italics* denote scalar variables; and the Helvetica font denotes literals. I also distinguish between types and tokens by using bold face for the former and plain font for the latter.

References

A. Abeillé, Y. Schabes, & A. K. Joshi. (1990) "Using Lexicalized Tree Adjoining Grammars for Machine Translation," *13th International Conference on Computational Linguistics*. Helsinki, Finland.

L. Ahrenberg, M. Andersson, & M. Merkel. (1998) "A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts," *36th Annual Meeting of the Association for Computational Linguistics*. Montreal, Canada.

R. K. Ahuja, T. L. Magnati, & J. B. Orlin. (1993) *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Englewood Cliffs, NJ.

Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, I. D. Melamed, F. J. Och, D. Purdy, N. A. Smith, & D. Yarowsky. (1999) "Statistical Machine Translation," CLSP Technical Report, Baltimore, MD. Available from *www.clsp.jhu.edu/ws99/projects/mt/final_report/mt-final-report.ps*

S. Aster. (1997) Personal communication.

R. H. Baayen & R. Lieber. (1997) "Word Frequency Distributions and Lexical Semantics," *Computers and the Humanities 30*, pp. 281–291.

Y. Bar-Hillel. (1964) *Language and Information*. Addison-Wesley, Reading, MA.

A. Blum & T. Mitchell. (1998) "Combining Labeled and Unlabeled Data with Co-Training," *11th Annual Conference on Computational Learning Theory*, Madison, WI, pp. 92-100.

R. Bellman. (1957) *Dynamic Programming*. Princeton University Press, Princeton, NJ.

J. Brousseau, C. Drouin, G. Foster, P. Isabelle, R. Kuhn, Y. Normandin, & P. Plamondon. (1995) "French Speech Recognition in an Automatic Dictation System for Translators: The TransTalk Project," *EuroSpeech'95*. Madrid, Spain.

P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, R. L. Mercer, & P. Roossin. (1988) "A Statistical Approach to Language Translation," *12th International Conference on Computational Linguistics*. Budapest, Hungary.

P. F. Brown, J. C. Lai, & R. L. Mercer. (1991a) "Aligning Sentences in Parallel Corpora," *29th Annual Meeting of the Association for Computational Linguistics*. Berkeley, CA.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, & R. L. Mercer. (1991b) "Word Sense Disambiguation Using Statistical Methods," *29th Annual Meeting of the Association for Computational Linguistics*. Berkeley, CA.

P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, & R. L. Mercer. (1992) "Class-Based *n*-gram Models of Natural Language," *Computational Linguistics 18*(4). 467–479.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, M. J. Goldsmith, J. Hajic, R. L. Mercer & S. Mohanty. (1993a) "But Dictionaries Are Data Too," *ARPA Workshop on Human Language Technology*. Princeton, NJ.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, & R. L. Mercer. (1993b) "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics 19*(2). 263–311.

C. Buckley. (1993) "The Importance of Proper Weighting Methods," *DARPA Workshop on Human Language Technology*. Princeton, NJ.

M.-H. Candito. (1998) "Building Parallel LTAG for French and Italian," *17th International Conference on Computational Linguistics.* Montreal, Canada.

R. Catizone, G. Russell, & S. Warwick. (1989) "Deriving Translation Data from Bilingual Texts," *First International Lexical Acquisition Workshop*. Detroit, MI.

H.-H. Chen, S.-J. Huang, Y.-W. Ding, & S.-C. Tsai. (1998) "Proper Name Translation in Cross-Language Information Retrieval," *36th Annual Meeting of the Association for Computational Linguistics*. Montreal, Canada.

S. Chen. (1993) "Aligning Sentence in Bilingual Corpora Using Lexical Information," *31st Annual Meeting of the Association for Computational Linguistics*. Columbus, OH.

K. W. Church, I. Dagan, W. Gale, P. Fung, J. Helfman, & B. Satish. (1993). "Aligning Parallel Texts: Do Methods Developed for English-French Generalize to Asian Languages?" *PacfoCol'93*. Taipei, Taiwan.

K. W. Church & E. H. Hovy. (1993) "Good Applications for Crummy Machine Translation," *Machine Translation 8*, 239–258.

K. W. Church. (1993) "Char_align: A Program for Aligning Parallel Texts at the Character Level," *31st Annual Meeting of the Association for Computational Linguistics*. Columbus, OH.

P. H. Cousin, L. Sinclair, J. F. Allain, & C. E. Love. (1990) *The Harper Collins French Dictionary*. Harper Collins Publishers, New York, NY.

P. H. Cousin, L. Sinclair, J. F. Allain, & C. E. Love. (1991) *The Collins Paperback French Dictionary*. Harper Collins Publishers, Glasgow.

T. M. Cover & J. A. Thomas. (1991) *Elements of Information Theory*. John Wiley & Sons, New York, NY.

I. Dagan, A. Itai, & U. Schwall. (1991) "Two Languages Are More Informative than One," *29th Annual Meeting of the Association for Computational Linguistics*. Berkeley, CA.

I. Dagan, S. Marcus & S. Markovitch. (1993a) "Contextual Word Similarity and Estimation from Sparse Data," *31st Annual Meeting of the Association for Computational Linguistics*. Columbus, OH.

I. Dagan, K. Church, & W. Gale. (1993b) "Robust Word Alignment for Machine Aided Translation," *Workshop on Very Large Corpora: Academic and Industrial Perspectives*. Columbus, OH.

I. Dagan & K. Church. (1994) "TERMIGHT: Identifying and Translating Technical Terminology," *Fourth ACL Conference on Applied Natural Language Processing*. Stuttgart, Germany.

I. Dagan. (1997) Personal communication.

B. Daille, É. Gaussier, & J.-M. Langé. (1994) "Towards Automatic Extraction of Monolingual and Bilingual Terminology," *15th International Conference on Computational Linguistics.* Kyoto, Japan.

F. Debili & E. Sammouda. (1992) "Appariement des Phrases de Textes Bilingues," *14th International Conference on Computational Linguistics.* Nantes, France.

W. E. Deming & F. F. Stephan. (1940) "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known," *Annals of Mathematical Statistics 11*, pp. 427-444.

A. P. Dempster, N. M. Laird, & D. B. Rubin. (1977) "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society 39(B)*, 1–38.

L. R. Dice. (1945) "Measures of the Amount of Ecologic Association Between Species," *Journal of Ecology 26*, pp. 297-302.

B. J. Dorr. (1992) "The Use of Lexical Semantics in Interlingual Machine Translation," *Machine Translation 7(3)*, pp. 135–193.

T. Dunning. (1993) "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics 19*(1), 61–74.

D. Elworthy. (1998) "Language Identification with Confidence Limits," *Sixth Workshop on Very Large Corpora*. Montreal, Canada.

D. A. Evans & C. Zhai. (1996) "Noun-Phrase Analysis in Unrestricted Text for Information Retrieval," *34th Annual Meeting of the Association for Computational Linguistics*. Santa Cruz, CA.

G. Foster, P. Isabelle, & P. Plamondon. (1996) "Word Completion: A First Step Toward Target-Text Mediated IMT," *16th International Conference on Computational Linguistics*. Copenhagen, Denmark.

P. Fung. (1995a) "Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus," *Third Workshop on Very Large Corpora*. Boston, MA.

P. Fung. (1995b) "A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora," *33rd Annual Meeting of the Association for Computational Linguistics*. Boston, MA.

P. Fung and K. W. Church. (1994). "K-vec: A New Approach for Aligning Parallel Texts," *15th International Conference on Computational Linguistics.* Kyoto, Japan. 1096-1102.

P. Fung and K. McKeown. (1994). "Aligning Noisy Parallel Corpora across Language Groups: Word Pair Feature Matching by Dynamic Time Warping," *1st Conference of the Association for Machine Translation in the Americas.* Columbia, MD. 81-88.

W. Gale & K. W. Church. (1991a) "A Program for Aligning Sentences in Bilingual Corpora," *29th Annual Meeting of the Association for Computational Linguistics*. Berkeley, CA.

W. Gale & K. W. Church. (1991b) "Identifying Word Correspondences in Parallel Texts," *DARPA Speech and Natural Language Workshop*. Asilomar, CA.

W. A. Gale & G. Sampson. (1995) "Good-Turing Frequency Estimation Without Tears" *Journal of Quantitative Linguistics 2*, pp. 217-237. Swets & Zeitlinger Publishers, Sassenheim, The Netherlands.

D. Graff, I. D. Melamed, & P. Morgovsky. (1997) "Hansard Corpus: Parallel Text in English and French" on CD-ROM, Linguistic Data Consortium.

B. Harris. (1988) "Bi-Text, a New Concept in Translation Theory," *Language Monthly #54.*

D. Hiemstra. (1996) *Using Statistical Methods to Create a Bilingual Dictionary*, Master's Thesis, University of Twente, The Netherlands.

D. Hiemstra. (1998) "Multilingual Domain Modeling in Twenty-One: Automatic Creation of a Bi-directional Translation Lexicon from a Parallel Corpus," *8th Meeting of Computational Linguistics in the Netherlands (CLIN).*

J. W. Hunt & T. G. Szymanski. (1977) "A Fast Algorithm for Computing Longest Common Subsequences," *Communications of the ACM 20(5)*, pp. 350-353.

P. Isabelle. (1992). "Bi-Textual Aids for Translators," *8th Annual Conference of the UW Centre for the New OED and Text Research*. Waterloo, Canada. 1-15.

P. Isabell. (1995). Personal communication.

M. Kay & M. Röscheisen. (1993) "Text-Translation Alignment," *Computational Linguistics 19*(1).

G. Kikui. (1998). "Term-list Translation Using Mono-lingual Word Co-occurrence Vectors," *36th Annual Meeting of the Association for Computational Linguistics*. Montreal, Canada.

A. Kilgarriff. (1997a) "I Don't Believe in Word Senses," *Computers and the Humanities 31*(2), 91–113.

A. Kilgarriff. (1997b) "What is Word Sense Disambiguation Good For?" *NLP Pacific Rim Symposium '97*. Phuket, Thailand.

K. Kita, T. Omoto, Y. Yano, & Y. Kato. (1993) "Application of Corpora to Second Language Learning: The Problem of Collocational Knowledge Acquisition," *2nd Workshop on Very Large Corpora*. Columbus, OH.

K. Knight & J. Graehl. (1997) "Machine Transliteration," *35th Annual Meeting of the Association for Computational Linguistics*. Madrid, Spain.

A. Kumano & H. Hirakawa. (1994) "Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistical Information," *15th International Conference on Computational Linguistics*. Kyoto, Japan.

J. Kupiec. (1993) "An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora," *31st Annual Meeting of the Association for Computational Linguistics*. Columbus, OH.

P. Langlais, M. Simard, & J. Véronis. (1998) "Methods and Practical Issues in Evaluating Alignment Techniques," *36th Annual Meeting of the Association for Computational Linguistics*. Montreal, Canada.

L. Langlois. (1996) "Bilingual Concordances: A New Tool for Bilingual Lexicographers," *2nd Conference of the Association for Machine Translation in the Americas*. Montreal, Canada.

H. Li & N. Abe. (1996) "Clustering Words with the MDL Principle," *16th International Conference on Computational Linguistics*. Copenhagen, Denmark.

H. Li & N. Abe. (1996) "Word Clustering and Disambiguation Based on Co-occurrence Data," *Proceedings of the 17th International Conference on Computational Linguistics*. Montreal, Canada.

J. M. Lucassen & R. L. Mercer. (1984) "An Information-Theoretic Approach to the Automatic Determination of Phonemic Baseforms," *IEEE International Conference on Acoustics, Speech and Signal Processing*. San Diego, CA.

E. Macklovitch. (1994) "Using Bi-textual Alignment for Translation Validation: The TransCheck System," *1st Conference of the Association for Machine Translation in the Americas*. Columbia, MD.

E. Macklovitch. (1995) "Peut-on verifier automatiquement la coherence terminologique?" *IV^{es} Journées scientifiques, Lexicommatique et Dictionnairiques,* organized by AUPELF-UREF. Lyon, France.

M. P. Marcus, B. Santorini, & M. A. Marcinkiewicz. (1993) "Building a Large Annotated Corpus of English: The Penn Treebank," *Computational Linguistics 19*(2).

J. S. McCarley. (1999) "Should We Translate the Documents or the Queries in Cross-Language Information Retrieval?" *37th Annual Meeting of the Association for Computational Linguistics*. College Park, MD.

T. McEnery & M. Oakes. (1995) "Cognate Extraction in the CRATER Project: Methods and Assessment," *From Texts to Tags: Issues in Multilingual Language Analysis, SIGDAT Workshop*. Dublin, Ireland.

I. D. Melamed. (1995) "Automatic Evaluation and Uniform Filter Cascades for Inducing *N*-best Translation Lexicons," *Third Workshop on Very Large Corpora*. Cambridge, MA.

I. D. Melamed. (1996a) "Automatic Construction of Clean Broad-Coverage Translation Lexicons," *2nd Conference of the Association for Machine Translation in the Americas*. Montreal, Canada.

I. D. Melamed. (1996b) "Porting SIMR to New Language Pairs," IRCS Technical Report 96-26. University of Pennsylvania, Philadelphia, PA.

I. D. Melamed. (1997a) "Measuring Semantic Entropy," *SIGLEX Workshop on Tagging Text with Lexical Semantics*. Washington, DC.

I. D. Melamed. (1997b) "A Portable Algorithm for Mapping Bitext Correspondence," *35th Conference of the Association for Computational Linguistics*. Madrid, Spain.

I. D. Melamed. (1997c) "A Word-to-Word Model of Translational Equivalence," *35th Conference of the Association for Computational Linguistics*. Madrid, Spain.

I. D. Melamed. (1997d) "Automatic Discovery of Non-Compositional Compounds," *Second Conference on Empirical Methods in Natural Language Processing*. Providence, RI.

G. A. Miller (ed.). (1990) *WordNet: An On-Line Lexical Database*. Special issue of *International Journal of Lexicography 4*(3).

J. Nerbonne, L. Karttunen, E. Paskaleva, G. Proszeky, & T. Roosmaa. (1997) "Reading More into Foreign Languages," *5th ACL Conference on Applied Natural Language Processing*. Washington, DC.

D. W. Oard (1997) "Adaptive Filtering of Multilingual Document Streams," *5th RIAO Conference*. Montreal, Canada.

F. J. Och. (1999) "An Efficient Method for Determining Bilingual Word Classes," *15th Annual Meeting of the European Association for Computational Linguistics*. Bergen, Norway.

H. Papageorgiou, L. Cranias, & S. Piperidis. (1994) "Automatic Alignment in Parallel Corpora," *32nd Annual Meeting of the Association for Computational Linguistics (Student Session)*. Las Cruces, NM.

F. Pereira, N. Tishby, & L. Lee. (1993) "Distributional Clustering of English Words," *31st Annual Meeting of the Association for Computational Linguistics*. Columbus, OH.

C. Phillips & T. J. Warnow. (1996) "The Asymmetric Median Tree—A New Model for Building Consensus Trees," *Discrete Applied Mathematics 71(1-3)*, pp. 331-335.

R. Rapp. (1995) "Identifying Word Translations in Non-Parallel Texts," Student Session, *33rd Annual Meeting of the Association for Computational Linguistics*. Boston, MA.

P. Resnik. (1997) "Evaluating Multilingual Gisting of Web Pages," *AAAI Symposium on Cross-Language Text and Speech Retrieval*. Stanford University, Stanford, CA.

P. Resnik & I. D. Melamed. (1997) "Semi-Automatic Acquisition of Domain-Specific Translation Lexicons," *5th ACL Conference on Applied Natural Language Processing*. Washington, DC.

P. Resnik, M. B. Olsen, & M. Diab. (1997) "Creating a Parallel Corpus from the Book of 2000 Tongues," *10th TEI User Conference*. Providence, RI.

P. Resnik & D. Yarowsky. (1997) "A Perspective on Word Sense Disambiguation Methods and Their Evaluation," *SIGLEX Workshop on Tagging Text with Lexical Semantics*. Washington, DC.

P. Resnik. (1999) "Mining the Web for Bilingual Text," *37th Annual Meeting of the Association for Computational Linguistics*. College Park, MD.

P. Resnik & T. Kanungo. (1999). Personal communication.

K. Ries, F. D. Buo, & A. Waibel. (1996) "Class Phrase Models for Language Modeling," *Fourth International Conference on Spoken Language Processing*. Philadelphia, PA.

L. Rüschendorf. (1995) "Convergence of the Iterative Proportional Fitting Procedure," *Annals of Statistics 23(4)*, 1160-1174.

V. Sadler & R. Vendelmans. (1990) "Pilot Implementation of a Bilingual Knowledge Bank," *13th International Conference on Computational Linguistics*. Helsinki, Finland.

H. Schuetze & J. O. Pedersen. (1995) "Information Retrieval Based on Word Senses," *Fourth Annual Symposium on Document Analysis and Information Retrieval*, pp. 161-175. Las Vegas, NV.

H. Schuetze. (1998) "Automatic Word Sense Discrimination," *Computational Linguistics 24*(1), pp. 97-124.

S. Shieber. (1994) "Restricting the Weak-Generative Capacity of Synchronous Tree-Adjoining Grammars," *Computational Intelligence 10:4*, pp. 371-385.

J. H. Shin, Y. S. Han, & K.-S. Choi. (1996) "Bilingual Knowledge Acquisition from Korean-English Parallel Corpus Using Alignment Method (Korean-English Alignment at Word and Phrase Level)," *16th International Conference on Computational Linguistics*. Copenhagen, Denmark.

M. Simard, G. F. Foster, & P. Isabelle. (1992) "Using Cognates to Align Sentences in Bilingual Corpora," *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*. Montreal, Canada.

M. Simard, G. F. Foster, & F. Perrault. (1993) "TransSearch: A Bilingual Concordance Tool." Centre d'Innovation en Technologies de l'Information, Laval, Canada.

M. Simard. (1995) Personal communication.

M. Simard & P. Plamondon. (1996) "Bilingual Sentence Alignment: Balancing Robustness and Accuracy," *2nd Conference of the Association for Machine Translation in the Americas*. Montreal, Canada.

F. Smadja. (1992) "How to Compile a Bilingual Collocational Lexicon Automatically," *AAAI Workshop on Statistically-Based NLP Techniques*. San Jose, CA.

N. A. Smith & M. E. Jahr. (2000) "Cairo: An Alignment Visualization Tool," *Second Conference on Language Resources and Evaluation*. Athens, Greece.

R. Sproat, C. Shih, W. Gale, & N. Chang. (1996) "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese," *Computational Linguistics 22*(3):377-404.

J. Svartvik. (1992) *Directions in Corpus Linguistics.* Mouton de Gruyter, Berlin.

D. Turcato. (1998) "Automatically Creating Bilingual Lexicons for Machine Translation from Bilingual Text," *36th Annual Meeting of the Association for Computational Linguistics*. Montreal, Canada.

R. V. V. Vidal. (1993) *Applied Simulated Annealing.* Springer-Verlag, Heidelberg, Germany.

S. Vogel, H. Ney, & C. Tillmann. (1996) "HMM-Based Word Alignment in Statistical Translation," *16th International Conference on Computational Linguistics*. Copenhagen, Denmark.

E. M. Voorhees. (1993) "Using Wordnet to Disambiguate Word Senses for Text Retrieval," *SIGIR '93*, pp. 171–180.

P. Vossen (ed.). (1998) *Eurowordnet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, the Netherlands.

S. Wan & M. Verspoor. (1998) "Automatic English-Chinese Name Transliteration for Development of Multilingual Resources," *36th Annual Meeting of the Association for Computational Linguistics*. Montreal, Canada.

Y. Wang, J. Lafferty, & A. Waibel. (1996) "Word Clustering with Parallel Spoken Language Corpora," *Fourth International Conference on Spoken Language Processing*. Philadelphia, PA.

W. Weaver. (1955) "Translation." In William N. Locke and Donald A. Booth, eds., *Machine Translation of Languages*. MIT Press, Cambridge.

J. S. White & T. A. O'Connell. (1993) "Evaluation of Machine Translation," *ARPA Workshop on Human Language Technology*. Princeton, NJ.

D. Wu. (1994) "Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria," *32nd Annual Meeting of the Association for Computational Linguistics*. Las Cruces, NM.

D. Wu & P. Fung. (1994) "Improving Chinese Tokenization with Linguistic Filters on Statistical Lexical Acquisition," *4th Conference on Applied Natural Language Processing*. Stuttgart, Germany.

D. Wu & X. Xia. (1994) "Learning an English-Chinese Lexicon from a Parallel Corpus," *First Conference of the Association for Machine Translation in the Americas*. Columbia, MD.

D. Wu. (1995) "Grammarless Extraction of Phrasal Translation Examples from Parallel Texts," *Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*. Leuven, Belgium.

D. Yarowsky. (1993) "One Sense Per Collocation," *DARPA Workshop on Human Language Technology*. Princeton, NJ.

D. Yarowsky. (1996) *Three Machine Learning Algorithms for Lexical Ambiguity Resolution*. Ph.D. Dissertation, University of Pennsylvania, Philadelphia, PA.

G. K. Zipf. (1936) *The Psycho-biology of Language: an Introduction to Dynamic Philology*. Routledge, London, UK.