Preface

Anyone who has worked with corpora will be all too aware of differences between them. Depending on the differences, it may, or may not, be reasonable to expect results based on one corpus to also be valid for another. It may, or may not, be appropriate for a grammar, or parser, based on one to perform well on another. It may, or may not, be straightforward to port an application from a domain of the first text type to a domain of the second. Currently, characterisations of corpora are mostly textual and informal. A corpus is described as "Wall Street Journal" or "transcripts of business meetings" or "foreign learners' essays (intermediate grade)". It would be desirable to be able to place a new corpus in relation to existing ones, and to be able to quantify similarities and differences.

Allied to corpus-similarity is corpus-homogeneity. An understanding of homogeneity is a prerequisite to a measure of similarity – it makes little sense to compare a corpus sampled across many genres, like the Brown, with a corpus of weather forecasts, without first accounting for the one being broad, the other narrow.

Given the centrality of corpora to contemporary language engineering, it is remarkable how little research there has been on corpus similarity. The only well-understood measure is cross-entropy, from Information Theory, which is widely used in language modelling, particularly for speech recognition (see, eg, Roukos 1996). However it is not clear whether, or where, it is a good measure, and there is some evidence that it does not match our intuitions (Kilgarriff and Rose 1998, Kilgarriff in press).

Biber's work (1989, 1995) on corpus characterisation, coming from sociolinguistics, has made a considerable impact, with various researchers applying the model in language engineering (eg Folch *et al* 2000) or subjecting it to critical scrutiny (Lee 2000). Studies in text classification, genre and sublanguage are also salient, but it is rarely evident how well the technologies developed in these fields are suited to measuring corpus similarity or homogeneity.

There are of course many ways in which two corpora will differ, and different kinds of difference will be relevant for different kinds of purposes. Thus, similarity such that a part-of-speech tagger developed for one corpus works well in the other, may differ from similarity for Machine Translation. We currently lack a sophisticated vocabulary for talking about the various ways in which corpora differ, and hope that the workshop will contribute to the development of one.

We welcomed contributions concerned with measuring and comparing corpora from any field.

References

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge University Press.

Biber, Douglas. 1995. *Dimensions in Register Variation*. Cambridge University Press.

Folch, Helka, Serge Heiden, Benôit Habert, Serge Fleury, Gabriel Illouz, Pierre Lafon, Julien Nioche and Sophie Prévost 2000. TyPTex: Inductive typological text classification by multivariate statistical analysis for NLP systems tuning/evaluation. In Proc. 2nd LREC, Athens, Greece. Pp 141–148.

Kilgarriff, Adam. In press. Comparing corpora. *Int. Jnl. Corpus Linguistics*.

Kilgarriff, Adam and Tony Rose. 1998. Measures for corpus similarity and homogeneity. In *Proc. EMNLP–3*, pages 46–52, Granada, Spain, June. ACL-SIGDAT.

Lee, David. 2000. *Modelling Variation in spoken and written language: the multidimensional approach revisited*. Ph.D. thesis, University of Lancaster.

Roukos, Salim, 1996. *Language Representation*, chapter 1.6. NSF and EU Survey of the State of the Art in Human Language Technologies, `www.cse.ogi/CSLU/HLTsurvey.html`.

Adam Kilgarriff
Co-chair

Programme Committee

Douglas Biber	Northern Arizona University
Jeremy Clear	University of Birmingham
Ted Dunning	MusicMatch Software, Inc.
Tomaž Erjavec	Jozef Stefan Institute, Slovenia
Pascale Fung	University of Science and Technology Hong Kong
Gregory Grefenstette	Xerox Research Centre Europe
Benôit Habert	LIMSI, France
Przemek Kaszubski	Adam Mickiewicz University, Poland
Adam Kilgarriff (Co-chair)	ITRI, University of Brighton
David Lee	University of Lancaster
Oliver Mason	University of Birmingham
Doug Oard	University of Maryland
Tony Rose	Canon Research
Tony Berber Sardinha (Co-chair)	Catholic University of Sao Paulo, Brazil
George Tambouratzis	ILSP, Athens
Christopher Tribble	King's College, London University

CONFERENCE PROGRAM
TABLE OF CONTENTS