The Human Language Chorus Corpus (HULCC)

Contact author: John Elliott,
J.Elliott@lmu.ac.uk,
Computational Intelligence Research Group, School of Computing,
Leeds Metropolitan University,
Leeds, LS1 3HE, England

Co-Author Debbie Elliott,
debe@comp.leeds.ac.uk
Centre for Computer Analysis of Language and Speech, School of Computing, University of Leeds,
Leeds, Yorkshire, LS2 9JT England

Abstract

Many aspects of linguistic research, whatever their aims and objectives, are reliant on cross-language analysis for their results. In particular, any research into generic attributes, universals, or inter-language comparisons, requires samples of languages in a readily accessible format, which are 'clean' and of adequate size for statistical analysis. As computer-based corpus linguistics is still a relatively recent discipline, currently available corpora still suffer from a lack of breadth and conformity. Probably due in part to restrictions dictated by funding, many of the machine-readable resources publicly available are for English or one of the major Indo-European languages and, although this is often frustrating for researchers, it is understandable. An equally problematic aspect of inter-corpus analysis is the lack of agreement between annotation schemes: their format, constituent parts-of-speech, granularity and classification, even within a single language such as English.

The aim of HuLCC is to provide a corpus of sufficient size to expedite such inter-language analysis by incorporating languages from all the major language families, and in so doing, also incorporating all types of morphology and word order. Parts-of-speech classification and granularity will be consistent across all languages within the corpus and will conform more closely to the main parts-of-speech originally conceived by Dionysius Thrax than to the fine-grained systems used by the BNC[1] and LOB[2] corpora. This will then enable cross-language analysis without the need for cross-mappings between differing annotation systems, or for writing/adapting software each time a different language or corpus is analysed. It is also our intention to encode all text using Unicode to accommodate all script types with a single format, whether they traditionally use standard ASCII, extended ASCII or 16 bits.

An added feature will be the inclusion of a common text element, which will be translated across all languages to provide both useful translation data and a precise comparable thread for detailed linguistic analysis. Initially, it is planned to provide at least 20,000 words for each chosen language, as this amount of text exceeds the point where randomly generated text attains 100% bigram and trigram coverage (Elliott, J, 2002). This significantly contrasts statistically with the much lower percentages attained by natural languages and provides a statistical rationale for what is often a hotly debated point.

Finally, as all constituent language samples within HuLCC conform to the same format and mark-up, a single set of tools will accompany what will be a freely available corpus for the academic world, to facilitate basic analytical needs. This paper outlines the rationales and design criteria that will underlie the development and implementation of this corpus.

1. Introduction

In many ways, the current state of Corpus Linguistics is analogous to when the Railways were a fledgling engineering discipline. Each individual enterprise proceeded to develop their own designs based on their own priorities and interpretation of 'best practice'. Finally, when the individual developments, all of which were convinced that their product designs were the correct ones to adopt, had to be integrated into a single

[1] British National Corpus
[2] Lancaster-Oslo-Bergen

network for the system to truly serve the users' needs, it was suddenly apparent that many lines were incompatible as many had chosen different gauge tracks.

Mirroring such fragmentation are the existing structures, tools and annotation schemes that proliferate even within a single language's representation across developed machine-readable corpora. If a goal can be achieved by the interrogation of a single corpus, then this may well not be an issue, but for any research that depends on cross-language or corpus analysis, this inability to 'sing from the same sheet-music' is frustrating at best. An example of this is how the BNC and LOB corpora annotation schemes compare with respect to their overall granularity and labelling criteria for parts-of-speech in some of there base forms. With respect to their granularity, the BNC lists 67 tags and the LOB corpus lists a tag set of 121 parts-of-speech. Given that these are two of the most widely used English corpora, the previous issues mentioned are already clearly corroborated. Looking at a representative set of tags, which label the 'base' forms of the main parts-of-speech (common noun, adjective, adverb, verb, conjunction, article and preposition), only one tag (common noun) was found to agree in its annotation labelling. Both corpora have the necessary engineering criteria (rails) but once again the gauge (or system) used by each system, does not 'mesh' and 'travel' across these resources without significant re-engineering.

Our aim is therefore to address such inconsistencies to help expedite comparative linguistics, whilst providing a resource that is readily usable for multi-lingual analysis.

2. Texts

The text genre for version one, and the 'core' resource of the corpus, will target 'everyday' usage of source languages in electronic format, rather than targeting either informative or imaginative prose in isolation. It is therefore the endeavour of the collation process to gather prose irrespective of subject matter but to avoid jargon-laden sources such as manuals or scientific papers. Where readily compiled and annotated sources are not obtainable, primary sources are likely to be newspaper or narrative-based to expedite text-type consistency and availability, whilst maintaining the contemporariness of the corpus.

3. Which languages

An initial and pragmatic proposal is to target the more widely used languages from each 'family': Indo-European, Indo-Iranian, Sin-Tibetan, Uralic, Altaic, Malayo-Polynesian (Austroesian), Niger-Congo, Tai, Austro-Asiatic, Afro-Asiatic, Amerindian (Eskimo-Aleut), Dravidian and some independents: it is worth noting that even the division of language families can be a contentious issue and the names used often differ, just as with parts-of-speech. This selection is intended to address both representative issues regarding population and language diversity. However, it is intended to later widen this initial selection of languages with priority on diversity irrespective of usage. Examples of the 'core' selection to comprise the human chorus will be: Mandarin, English, Hindi, Russian, Arabic, Malay, Japanese, Tamil, Greek, Navajo, Hungarian, Latin, French, Gaelic, Turkish, and Swahili.

Of course, the main problem to-date has been acquiring useful 'clean' electronically readable samples of some of these less accessible languages and their subsequent annotation. This single issue is far from trivial, where readily annotated resources are unavailable for inclusion: whether purchased or donated. All grammatical annotation for these corpora will be checked by human experts, so the time frame is likely to be an issue also.

Nevertheless, once the texts for each language sub-corpus have been acquired, with their particular 'local' annotation schemes, they will form the 'back-bone' of the corpus, from which subsequent meta-tagging, translation threads and tools will be added.

4. Structure

The corpus will initially comprise the selected languages as isolated sub-corpora to provide language specific or restricted cross-language analysis. In addition to this and to complement the main text resources that comprise the corpus, a 'common' text thread will also be included. One of the most readily available resources for such a task is the bible. However, as the corpus is intended to represent current 'everyday' language, the preference for a modern publication to represent the quasi-stationary behaviour of contemporary language for such multi-lingual translation becomes an overriding one. The choice is still to be made.

The translation thread will have an identical annotation mark-up scheme to the main body of the corpus, as this will then provide additional material for any statistical analysis, which can be merged with all other text for that language and avoid any redundancy. The annotated texts that constitute the translation thread will need to be aligned for the purposes of analysing mappings between syntax and semantic content. However, no conclusion has yet been reached concerning the segmentation level. Sentence alignment is almost a 'knee-jerk' response to this problem but this is far from 'fail-safe', as translators often have reason to merge or split sentences for the purposes of style and readability in the target language. Comparative statistical analyses of many languages at sentence level has shown that this segmentation device relies heavily on style, rather than on any 'true' linguistic imperatives, and can differ considerably from language to language, as well as from author to author. It is due to this issue that alignment at phrase level is currently under consideration; this level of segmentation exhibits 'true' and consistent language behaviour due to overriding cognitive constraints in language generation (Elliott, J. 2000 & Frederici, 2001).

5. Size

The sample size of a corpus is often a hotly debated topic, and answers to this particular question usually gravitate towards 'the bigger the better' (EAGLES, 1996) particularly with respect to issues of data sparseness and word prediction accuracy derived from inter-word statistics (Lesher, Moulton & Higginbottom, 1999), but just as often for pragmatic reasons. However, normal theoretical principles of statistical sampling and inference do not apply, as it is often impossible to delimit the total population in any rigorous way. There is also no obvious unit of language, which is to be sampled and which can be used to define the population (Atkins et al, 1992). For the purposes of getting version one of this corpus 'up and running', comparative behavioural characteristics of language are the primary objective.

It is with this aim in mind that initial plans are to provide at least 20,000 words for each chosen language, as this amount of text exceeds the point where randomly generated text attains 100% bigram and trigram coverage (Elliott, J. 2001) and natural language has generated all probable combinations occurring in 'common' usage. This point, where all possible combinations are generated at trigram segmentation, significantly contrasts statistically with the much lower percentages attained by natural languages. This particular metric was originally chosen for rationalising a minimum transmission length to transparently communicate natural language and to prevent any recipient from discarding it as a random event. To add further weight to this rationale, preliminary statistical analyses show that after approximately 14,000 words reliable scores can be obtained in machine translation output evaluations (when comparing several systems), and that any additional sampling only serves to confirm results (Elliott, D et al, 2003).

Sparseness is even an issue with the largest of project proposals. A billion word annotated corpus would be a huge undertaking, nevertheless, it is still a finite resource and would still display remarkably sparse data for most of its word list (Sinclair, 91): statistical observations encapsulated in the properties of a Zipfian distribution (Zipf, 1949). However, HuLLC's primary aim, for version one of its main body of text, is not to analyse individual word behaviour exhaustively, except for perhaps the higher-ranked content words and their relationship with function words, but to analyse inter-part-of-speech behaviour and morphological issues comparatively across languages. Subsequent versions will then increase word counts for all languages with the aim of providing a resource suitable for most analytical objectives.

6. Corpus mark-up

The grammatical mark-up of constituent languages within the corpus is a major issue, due to both the diversity of the tag-sets, composed for individual corpus aims, and linguistic interpretation. Even at the level of lexico-grammatical wordclass annotation (Part-of-Speech wordtagging), which corresponds to layers 'a' and 'b' outlined in the EAGLES report (EAGLES, 1996), there is a great diversity of schemes and models available. Here an example sentence is tagged according to several alternative tagging schemes and vertically aligned (Atwell et al, 1996):

Brown	ICE	LLC	LOB	PARTS	POW	SEC	UPenn	
select	VB	V(montr,imp)	VA+0	VB	adj	M	VB	VB
the	AT	ART(def)	TA	ATI	art	DD	ATI	DT
text	NN	N(com,sing)	NC	NN	noun	H	NN	NN
you	PPSS	PRON(pers)	RC	PP2	pron	HP	PP2	PRP
want	VB	V(montr,pres)	VA+0	VB	verb	M	VB	VBP
to	TO	PRTCL(to)	PD	TO	verb	I	TO	TO
protect	VB	V(montr,infin)	VA+0	VB	verb	M	VB	VB
.	.	PUNC(per)	.	.	.	.	.	.

Table 1

EAGLES layers of syntactic annotation:

(a) Bracketing of segments
(b) Labelling of segments
(c) Showing dependency relations
(d) Indicating functional labels
(e) Marking sub-classification of syntactic segments
(f) Deep or ‘logical’ information
(g) Information about the rank of a syntactic unit
(h) Special syntactic characteristics of spoken language

In Jan Cloeren’s paper, which evaluates schemes for a cross-linguistic tagset (Cloeren, 1993), the tagging of Germanic languages is considered in detail, with the following conclusion as its basic cross-linguistic tagset:

Noun
Pronoun
Article
Adjective
Numeral
Verb

Adverb
Preposition
Conjunction
Particle
Interjection
Formulaic expression

Erjavec, Ide and Tufis compare, by language, the number of attributes in each part of speech for six central and eastern European languages (Erjavec et al, 98). Their conclusions, tabulated below (Table 2), illustrate both granularity and, perhaps more importantly, the absence of features (denoted by hyphens) in individual languages in accordance with morpho-syntactic descriptions (MSDs) developed from proposals in the EAGLES project, with subsequent modifications for the Multi-East project. A zero indicates that it distinguishes no features for that part-of-speech. However, as illustrated in Table 3 below, interpretation of grammatical tokens suffers from a lack of classification universality and devices indicated as absent or rare in a language may well exist.

Chinese 1	Chinese 2	LOB	LOB	BNC	Cuban-Spanish	Bulgarian
AS	ACN	ABL:	NP	AJ0	ART.	Noun
CC	AJO	ABN:	NP$	AJC	VAR.PRON.	Prepos
CD	AJS	ABX:	NPL	AJS	V.IND.	Verb
CL	ASO	AP	NPL$	AT0	PREP.	Adj
DR	AUX	AP”	NPLS	AV0	SUST.M.SING.	Conj
DT	AVO	APS	NPLS$	AVP	PRON.PERS.	Num
IN:	CJO	AP$	NPS	AVQ	ADV.	Pron
JJ:	DE	APS$	NPS$	CJC	ADJ.DET.	Adv
LC	ELM	AT	NPT	CJS	SUST.M.PL.	Part
MD	FIX	ATI	NPT”	CJT	SUST.F.PL.	Interj
MJ:	IDM	BE	NPTS	CRD	ADJ.CALIF.P.	
MR	MC	BED	NPTS$	DPS	CONTRAC.	
NN	NMW	BEDZ	NR	DT0	PRON.INT-EXC	
NR:	NNC	BEG	NR$	DTQ	PRON.REL.	
NV	NND	BEM	NRS	EX0	SUST.DIMIN.	
PN	NNO:	BEN	NRS$	ITJ	ADJ.CALIF.A.	
RB	NPO	BER	OD	NN0	PER.V.	
SC	PND	BEZ	OD$	NN1	PRON.INDEF.	
VA	PRF	CC	PN	NN2	SUST.PR.-GEO	
VC	PRP	CC”	PN”	NP0	V.SUBJ.	
VE	SUF	CD	PN$	ORD	PRON.POS.	
VS	TIM	CD$	PN$”:	PNI	V.ENCL.	
WD	VVO	CD-CD	PP$:	PNP	LEX.C.-FECH.	
	XXO	CD1	PPS$	PNQ	PRON.DEMOST.	
		CD1$	PP1A	PNX	SGL.	
		CD1S	PP1A$	POS	SUST.AUMENT.	
		CDS	PP1O	PRF	INTERJ.	
		CS	PP1OS	PRP	V.IMP.	
		CS”	PP2	PUL	ADJ.DIMIN.	
		DO	PP3	PUN		
		DOD	PP3A	PUQ		
		DOZ	PP3AS	PUR		
		DT	PP3O	TO0		
		CT$	PP3OS	UNC		
		DTI	PPL	VBB		
		DTS	PPLS	VBD		
		DTX	PPLS”:	VBG		
		EX	QL	VBI		
		HV	QLP	VBN		
		HVD	RB	VBZ		
		HVG	RB”	VDB		
		HVN	RB$	VDD		
		HVZ	RBR	VDG		
		IN	RBT	VDI		
		IN”	R1	VDN		
		JJ	RN	VDZ		
		JJ”	RP	VHB		
		JJB	TO	VHD		
		JJB”	TO”	VHG		
		JJR	UH	VHI		
		JJR”	VB	VHN		
		JJT	VB”	VHZ		
		JJT”	VBD	VM0		
		JNP	VBG	VVB		
		MD	VBN	VVD		
		NC	VBZ	VVG		
		NN	WDT	VVI		
		NN”	WDT”	VVN		
		NN$	WDTR	VVZ		
		NNP	WP	XX0		
		NNP$	WP$	ZZ0		
		NNPS	WP$R			
		NNPS$	WPA			
		NNS	WPO			
		NNS”	WPOR			
		NNS$	WPR			
		NNU	WRB			
		NNU”	XNOT			
		NNUS	ZZ			

Table 2

	Romanian	Bulgarian	Czech	Slovene	Estonian	Hungarian
Noun	6	5	5	5	3	7
Verb	7	8	10	8	8	5
Adjective	7	3	7	5	3	8
Pronoun	8	8	12	10	4	7
Adverb	3	1	2	2	0	4
Adposition	4	1	3	3	1	1
Conjunction	5	2	3	2	1	3
Numeral	7	5	7	5	4	7
Interjection	0	1	0	0	0	1
Residual	0	0	0	0	0	0
Abbreviation	5	0	0	0	3	0
Particle	2	2	0	0	-	-
Determiner	8	-	-	-	-	-
Article	5	-	-	-	-	1

Table 3

This issue becomes crucial, when inheriting annotation schemes and expert lexico-grammatical wordclass annotation classification rationales, for the interpretation of what criteria constitute the allocation of a word-tag pairing: is a verb a verb in every language regardless of the fact that the word describes an action? To illustrate this point, the following words were entered into an online translator for single words entries of Thai-English to ascertain whether the parts-of-speech allocated agreed with the English interpretation.

Word	Thai PoS classification
Beautiful	V[8], N[1]
Run	V[4]
Man	N[5]
Sweet	V[1]
Old	N[3], V[5]
Tall	V[1]
Happy	V[5]
Blue	N[2]
White	V[2], N[1]
Fat	N[2]
Ugly	V[3]
Clever	N[1], V[1]
Quick	V[4]
Slow	V[4], N[1]
Wet	V[5]
Hot	V[5]
Cold	V[3], N[1]
Sexy	ADV[1]
Big	V[1]

Table 4

Results for these (see Table 4) predominantly adjectival words in the English language illustrate how classification can differ markedly, in addition to any labelling or granularity issues. This simple exercise demonstrates that any meta-tagging, in addition to providing a consistent, comparative baseline, will need to consider such differences in interpretation during the mapping process. A further issue is that of morphology: case markers, word-type determiners and the concatenation of lexical information in agglutinative languages to mention a few. These, of course, complicate matters further and will require a considerable investment of resources to assure consistent inter-language mapping across lexical elements, metaphors, clichés and word translation granularity for subsequent robust analysis.

An example of such mapping is illustrated here between two of the more closely related languages (English – French) e.g. He saw her duck

French:	Il	a	vu son	canard
	He	saw	her *(possessive)*	duck *(noun)*
	Il	l'	a vue	se baisser vivement
	He	her *(direct object)*	saw	duck *(verb)[lower herself quickly]*

This kind of sentence can easily be misinterpreted by a human, let alone a cross-linguistic or Machine Translation system.

It has been observed that there is a certain level of agreement between languages for such syntactic labelling. However, grammar is not indigenous to many languages such as Chinese, and the notion of parts-of-speech were most likely transplanted, and are a modified version of Western grammar, as originally devised by classical Mediterranean grammarians such as Pannini and Thrax.

Nevertheless, irrespective of these often-transplanted notions of grammar, the information we all communicate consists of the same physics and basic necessary building blocks to describe our environment and thought processes. As a human race, our mechanism for language processing and generation – the Brain – functions using the same physiology: areas of the Brain are dedicated to storing and accessing particular words classified by their parts-of-speech, such as the frontal lobe for verbs and the temporal lobe for verbs (Frederici, 2000). These neural constraints do not vary according to some linguistically geographical accident. So taking a theoretical stance akin to Chomsky, the *principles* should be detectable as long as the *parameters* are mapped accurately. This provides the rationale for such design criteria and the notion of a 'universal' base-set across which annotation can operate.

Assuming that our term 'universal' relates to fully developed mature scripts, the following ten features can be found (Aitchison, 1996).

All languages:

- Have consonants and vowels.
- Combine sounds into larger units.
- Have nouns – words for people and objects.
- Have verbs – words for actions.
- Can combine words.
- Can say who did what to whom.
- Can negate utterances.
- Can ask questions.
- Involve structure-dependence.
- Involve recursion.

The proposed format for annotation is a bracketed, hierarchical tagset, comprising 4 potential word classification layers: e.g. Corpus [N] {com; sg} <NN1> "open info"]

1. Generic base-set

eg. Noun

Wordclass: N

2. added information

Subclass:	com (common)	
	prop (proper)	
	number:	sg (singular)
		plu (plural)
	gender:	masc (masculine)
		fem (feminine)
		neut (neuter)
		case:
		nom (nominative)
		gen (genitive)
		dat (dative)
		acc (accusative)

3. original annotation scheme: scheme inherited from donated annotated corpus
4. open: additional information added by user

Morphological information is an important element of segmenting grammatical tokens, especially when mapping the syntactic content across agglutinative and inflectional languages to isolating and mixed morphologies. It is therefore intended to incorporate morphological information by concatenating grammatical tags, where words contain more than one grammatical element, to expedite content transparency and prevent misinterpretation of 'true' lexico-grammatical comparisons.

7. Text Formatting

As the perceived future of text encoding is seen as adopting Unicode, a 16 bit character set, which will negate the need for specialised fonts and mapping, used with current 8 bit formatting, the HuLCC corpus will future proof by converting all resources to this format prior to inclusion. Unicode also facilitates the accommodation of all script types within a single format, whether they traditionally use standard ASCII, extended ASCII or 16 bits encoding such as Big 5, so it is an ideal format for such a diversity of scripts.

This will have implications regarding the programming of analytical tools within the corpus, as most legacy tools are encoded for processing 8 bit formatting. However, investment in the development these bespoke analytical tools is essential for the corpus as a resource for extrapolating information and cross-linguistic analysis without further mapping issues.

8. Survey questionnaire

In the system design stage of the corpus, a questionnaire will be disseminated to all agreed participants for expert feedback on a range of design and implementation issues, with a view to facilitating the accommodation of as many issues as practicable. It is envisaged that the questionnaire will be mounted as an html document via our web site and any additional ad hoc participants will be most welcome. It is also planned to prototype the system at least once for further feedback, prior to the final implementation, to address additional issues such as HCI (Human Computer Interaction).

9. Basic tools

The array of tools planned for this corpus will, in addition to standard applications such as a concordancer and a suite of ngram analysis tools, include a visualisation toolkit for tasks such as bi-directional constraint clustering. The following tools are designed as language independent and suitable for unsupervised natural language learning:

- word-tag splitter for separating out words from their tags. This tool will facilitate separate part-of-speech and word level behavioural characteristics by presenting selected language texts filtered for subsequent analysis.

- Toolkit for visualising language structure and grammatical collocations patterns. Computational Linguistics researchers are showing increasing interest in Corpus-based Natural Language

Learning: empirical techniques to extract characteristic combinational and collocational patterns from Corpora or large-scale language datasets. Such grammatical collocation patterns can be used in systems for Part-of-Speech tagging of new, unseen text. Visualization profiles highlight distinguishing grammatical collocation behaviour between pairs of words or grammatical part-of-speech categories, allowing us to assess syntactic separation of related word-classes (Elliott, J. 2001). This tool is intended to enhance standard collocation analysis and incorporates statistical information of its own, with indicators for dependency.

- Concordancer: standard tool.

- Ngram suite: standard ngram statistical analysis at letter and word level.

- Cluster analysis: tools will be incorporated for clustering tokens, using a variety of metrics, both hierarchical and non-hierarchical. In addition to the more 'common' techniques, a tool using bi-directional constraint algorithms, based on content-functional word collocations is planned for inclusion: a recent research avenue, using functional words as constraints, where a single 'wildcard' word is constrained by two closed class - words [fwd_i <x> fwd_j] has been found to be a powerful clustering technique (Elliott, 2002b)

10. Corpus growth and feedback

Corpus growth is planned to be an ongoing undertaking, resulting in periodic increases in resources across all language groups: implementations of corpus growth will be evident by the numeric increment, e.g. 1.1, 1.2 etc. Post implementation of version one of the corpus will incorporate feedback taken from an online form permanently accessible to users. The feedback will then form part of the design phase of the next version.

11. Conclusions

Version one of this corpus will be designed principally for researchers looking into comparative linguistics and the search for computational universals. However, this is complemented by a translation thread incorporated throughout the constituent languages of the corpus, which is designed to facilitate analysis of directly comparable prose across a wide variety of linguistic structures, without the usual inherent annotation mapping problems. In the absence of a standard lexico-grammatical annotation model, the annotation mark-up rationale has been taken from prior computational analysis across many languages and cognitive constraints observed in neuroscience. Ultimately, the aim of this corpus is to represent the human language chorus by incorporating a set of languages that embody all morpho-syntactic mechanisms currently existing, within a single a coherent model, for accessible and robust analysis and the understanding of language structure.

References

Aitchison J. 1996 *The Seeds of Speech: Language Origin and Evolution*. Cambridge University Press, Cambridge.

Atkins S, Clear J H and Ostler N. 1992 Corpus Design Criteria in *Literary and Linguistic Computing*, Vol. 7, No. 1, pp. 1-16.

Atwell E, Demetriou G, Hughes J, Schiffrin A, Souter C, and Wilcock S. 2000 A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal*, volume 24, pages 7-23, International Computer Archive of Modern and medieval English, Bergen. ISSN: 0801-5775

Atwell E 1996 *Comparative Evaluation of Grammatical Annotation Models* in Sutcliffe R, Koch H-D, and McElligott A (editors), Industrial Parsing of Technical Manuals, pages 25-46, Rodopi, Amsterdam. ISBN: 90-420-0114-3 (hardback), 90-420-0102-X (paperback).

Cloeren, Jan 1993 "Towards a cross-linguistic tagset", *Proceedings of the ACL Workshop on Very Large Corpora*, Ohio State University, Columbus (OH), 1993.

EAGLES 1996 WWW site for European Advisory Group on Language Engineering Standards, http://www.ilc.pi.cnr.it/EAGLES96/home.html Specifically: Leech, G, R. Barnett and P. Kahrel, *EAGLES Final Report and guidelines for the syntactic annotation of corpora*, EAGLES Report EAG-TCWG-SASG/1.5.

Elliott, John. 2002a The Filtration of Inter-Galactic Objets Trouvés and the Identification of the Lingua ex Machina Hierarchy in: Proceedings of *World Space Congress*: *The 53rd International Astronautical Congress*, pp. 9.2.10. 2002, Huston, Texas, USA.

Elliott, John 2002b Detecting Languageness: in proceedings of *6th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI2002)*, Orlando, Florida, USA: volume XI, pp 323-328.

Elliott, J & Atwell, E. 2001 Visualisation of long distance grammatical collocation patterns in language in: IV2001: *Proceedings of 5th International Conference on Information Visualisation*, pp.297-302. 2001. ISBN 0-7695-195-

Elliott, J. Atwell, E & Whyte, B. 2000 Language identification in unknown signals: in Proceeding of *COLING'2000, 18th International Conference on Computational Linguistics*, pages 1021-1026, Association for Computational Linguistics (ACL) and Morgan Kaufmann Publishers, San Francisco. ISBN: 1-55860-717-X (2 volumes).

Elliott, D. Hartley, A. & Atwell, A. 2003 Rationale for a multilingual corpus for machine translation evaluation: In proceedings of *Corpus Linguistics 2003*, University of Lancaster.

Erjavec T, Ide N & Tufis D. 1998 Development And Assessment Of Common Lexical Specifications For Six Central And Eastern European Languages. *LREC'98.*

Friederici, A. D., Hickok, G. & Swinney, D. (eds.) 2001. Brain Imaging and sentence processing. Special Issue, *Journal of Psycholinguistic Research, Volume 30. New York: Kluwer Academic/Plenum Publishers.*

Lesher G W, Moulton B J & Higginbotham D J.1999. Effects of ngram order and training text size on word prediction. *Proceedings of the RESNA '99 Annual Conference*, 52-54, Arlington, VA: RESNA Press.

Piao Scott Songlin, 2000. Sentence and Word Alignment between Chinese and English (PhD Thesis), Lancaster University. (b): Piao, Scott Songlin ,2000. Chinese Corpus adapted from CEPC Corpus, Sheffield University, Sheffield UK.

Sinclair, John. 1991 *Corpus Concordance Collocation*. Describing English Language. Oxford: Oxford University Press.

Zipf, G. K. 1949 *Human Behaviour and The Principle of Least Effort*, Addison Wesley Press, New York.