ХҮМҮҮНЛЭГИЙН УХААНЫ ИХ СУРГУУЛЬ

Corpus-Based Language Studies

Munkhtuya Buyandelger (PhD)

Introduction

- Corpus linguistics: past and present
- What is corpus?
- Why use computers to study language?
- The corpus-based approach vs. the intuition-based approach
- Corpus linguistics: a methodology or theory?
- Case study: Using a language corpus for constructing a dictionary for specific purposes
- Summary

Corpus linguistics: Past and present

- However, corpus linguistics first appeared only in the early 1980s (cf.Leech 1992: 105), corpus-based language study has a rich history.
- The corpus methodology traces back to the pre-Chomskyan period when it was used by the field linguists such as Boas (1940) and linguists of the structuralist tradition, including Sapir, Newman, Bloomfield and Pike (see Biber and Finegan 1991:27)
- As McEnery and Wilson (2001:2-4) note, the basic corpus methodology was widespread in linguistics in the early 20th century.

1940-1970s Pre-Chomskyan era	1980's Chomskyan era	Modern era
-Field linguists (Boas, 1940) -Linguists of the structuralist tradition (Sapir, Newman, Bloomfield, Pike) -shoeboxes filled with paper slips rather than computers as a means of data storage -corpus methodology was severely criticized so that it became marginalized because of "skewness" of corpora. (Chomsky 1962; McEnery and Wilson 2001:5-13) (See Biber and Finegan 1991:207)	-Thanks to the invention of a computer and , corpus linguistics first appeared only in the early 1980's (cf.Leech 1992:105) -	-McEnery and Wilson (2001:2-4)note, the basic corpus methodology was widespread in linguistics in the early twentieth century. -With the development of more powerful computers offering ever increasing processing power and massive storage at relatively low cost, the exploitation of massive corpora became feasible. -The marriage of corpora with computer technology rekindled interest in the corpus methodology.

Examples of Language Corpuses

- **The Brown Corpus**- the Brown University Standard Corpus of Present day American was compiled in the 1960s by Henry Kučera and W. Nelson Francis at Brown University, Providence, Rhode Island as a general corpus (text collection) in the field of corpus linguistics. It contains 500 samples of English-language text, totaling roughly one million words, compiled from works published in the United States in 1961.
- **British National Corpus** (BNC)- The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English, both spoken and written, from the late twentieth century. http://www.natcorp.ox.ac.uk/
- **The Lancaster-Oslo/Bergen Corpus (LOB)** was compiled by researchers in Lancaster, Oslo and Bergen. It consists of one million words of British English texts from 1961. The texts for the corpus were sampled from 15 different text categories. Each text is just over 2,000 words long (longer texts have been cut at the first sentence boundary after 2,000 words) and the number of texts in each category varies (see table below). Further information about the texts can be found in the LOB manual (external link).

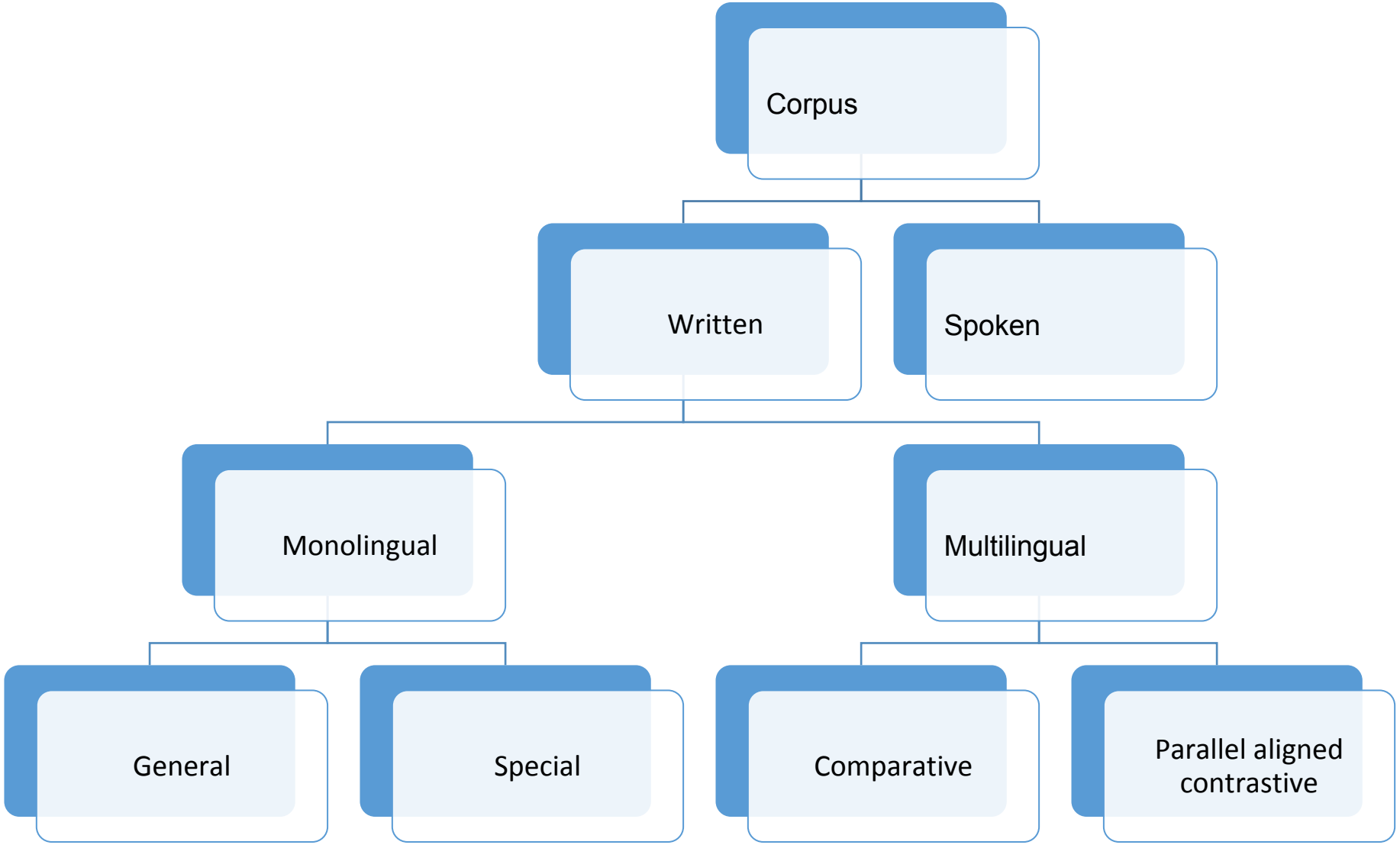
What is corpus?

Linguistics . a body of utterances, as words or sentences, assumed to be representative of and used for lexical, grammatical, or other linguistic analysis.

In modern linguistics, a corpus can be defined as a body of naturally occurring language, though strictly speaking:

> It should be added that computer corpora rarely haphazard collections of textual material: They are generally assembled with particular purposes in mind, and are often assembled to be (informally speaking) ***representative*** of some language or text type.
>
> (Leech 1992: 116)

Corpus types and classification



Main properties for corpora

‘A corpus is a collection of pieces of language that are selected and ordered according to ***explicit linguistic criteria*** in order to be used as a sample of the language.’

The ***‘linguistic criteria’,*** which are external to the texts themselves and dependent upon the intended use for the corpus (see Aston and Burnard 1998: 23), are used to select and put together these texts ‘in a principled way’(Johansson 1998: 3)

Main properties for a corpus

There is an increasing consensus that corpus is a collection of

1. ***in machine-readable*** / electronic files (may be annotated with various forms of linguistic information)
2. ***authentic texts*** (including transcripts of spoken data) which is
3. ***sampled*** texts to be
4. ***representative*** of a particular language or language variety (written or spoken).

Why use computers to study language?

A corpus include machine-readability, authenticity and representativeness.

1. **Authenticity** and **Representativeness** - enables every single data in reality
2. **Machine-readability** is a *de facto* attribute of modern corpora. Electronic corpora have advantages unavailable to their paper-based equivalents. The most obvious advantage of using a computer for language study is the **speed of processing** it affords
3. **ease with which it can manipulate data** (e.g. searching, selecting, sorting and formatting)

Advantages of computer-assisted language study

1. Computerized corpora can be processed and manipulated rapidly at minimal cost.
2. Computers can process machine-readable data accurately and consistently (see Barnbrook 1996: 11)
3. Computers can avoid human bias in an analysis, thus making the findings more reliable.
4. machine-readability allows further automatic processing to be performed on the corpus so that corpus texts can be enriched with various metadata and linguistic analyses.

Finally, it is the use of computerized corpora, together with computer programs which facilitate linguistic analysis, that distinguishes modern machine-readable corpora from early 'drawer-cum-slip' corpora.

(Svartvik 1992: 9)

The corpus-based approach vs. the intuition-based approach

Intuition-based approach

-can invent purer examples instantly for analysis
- can be influenced by one's dialect or sociolect
-cannot include all the typical representativeness in a language

(*Seuren 1998: 260-262*)

Corpus-based approach

-draws upon authentic and or real texts (*Leech 1991:14*)
-can include all the varieties of language in large quantity
-can yield reliable and improved quantitative data based on empirical data.

(*Francis, Hunston and Manning 1996; Chief Hung, Chen, Tsai and Chang 2000*)

Corpus linguistics: a methodology or theory?

Corpus linguistics (CL) is a methodology rather than an independent branch of linguistics. This view, however, is not shared by all scholars.

For example, it has been argued that corpus linguistics 'goes well beyond this methodological role' and has become an independent 'discipline'. (Tognini-Bonelli 2001: 1).

While we agree that CL is 'really a domain of research' and has become a new research enterprise and a new philosophical approach to linguistic enquiry', we maintain that CL is indeed a methodology rather than an independent branch of linguistics in the same sense as phonetics, syntax, semantics and pragmatics.

Case study: Using available corpora in reality

Title of case study: Compiling a dictionary for specific purposes using a language corpus

Aiming for establishing a corpus including technical terms and terminologies, On-board diagnostic codes used in automobile sector

Theoretical perspectives: interdisciplinary-Corpus linguistics and lexicography

Research methods: Three-A models of CL, corpus-based approaches, text /genre analysis, gathering statistics, focus group interview etc.

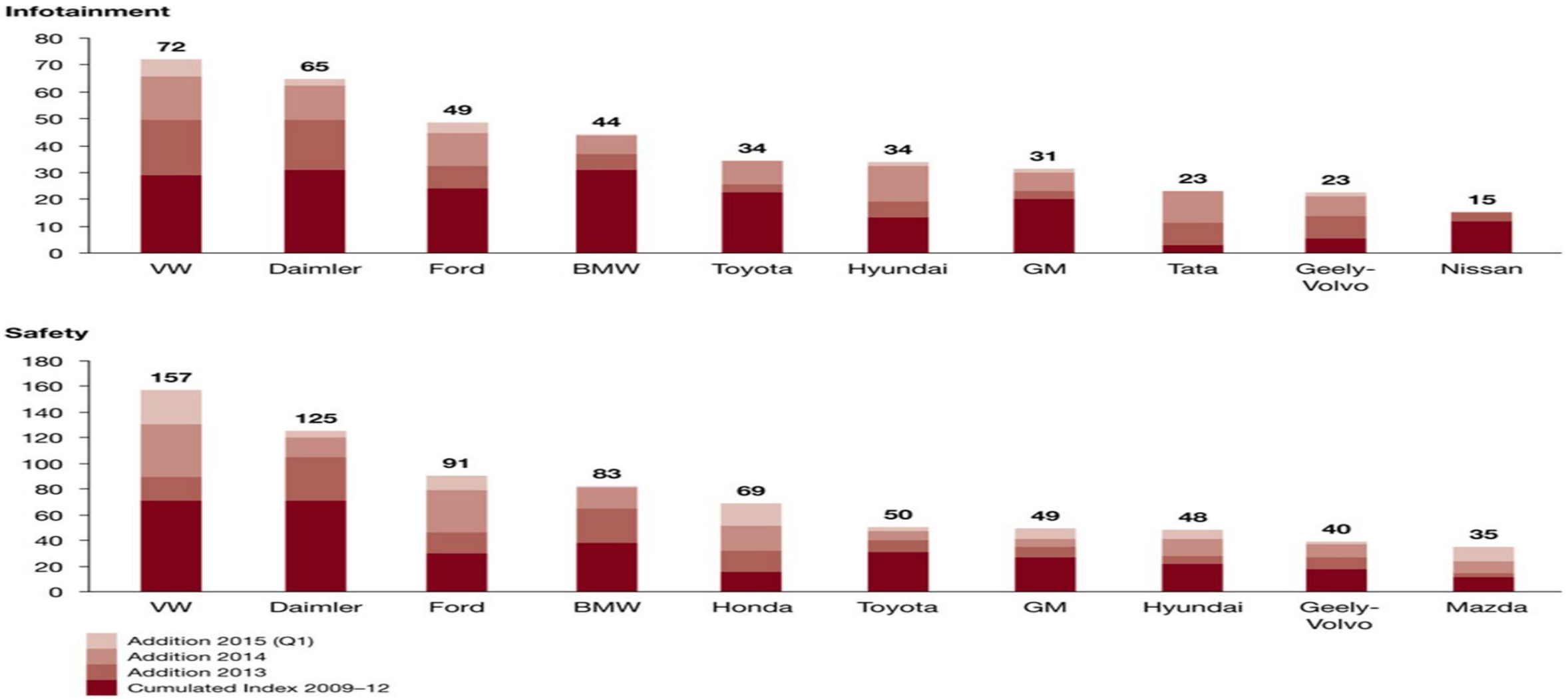
Available corpus: British National Corpus, internet corpus

(https://www.motorera.com/dictionary/)

Car innovation activity by automakers (2009-2015)

Exhibit 3

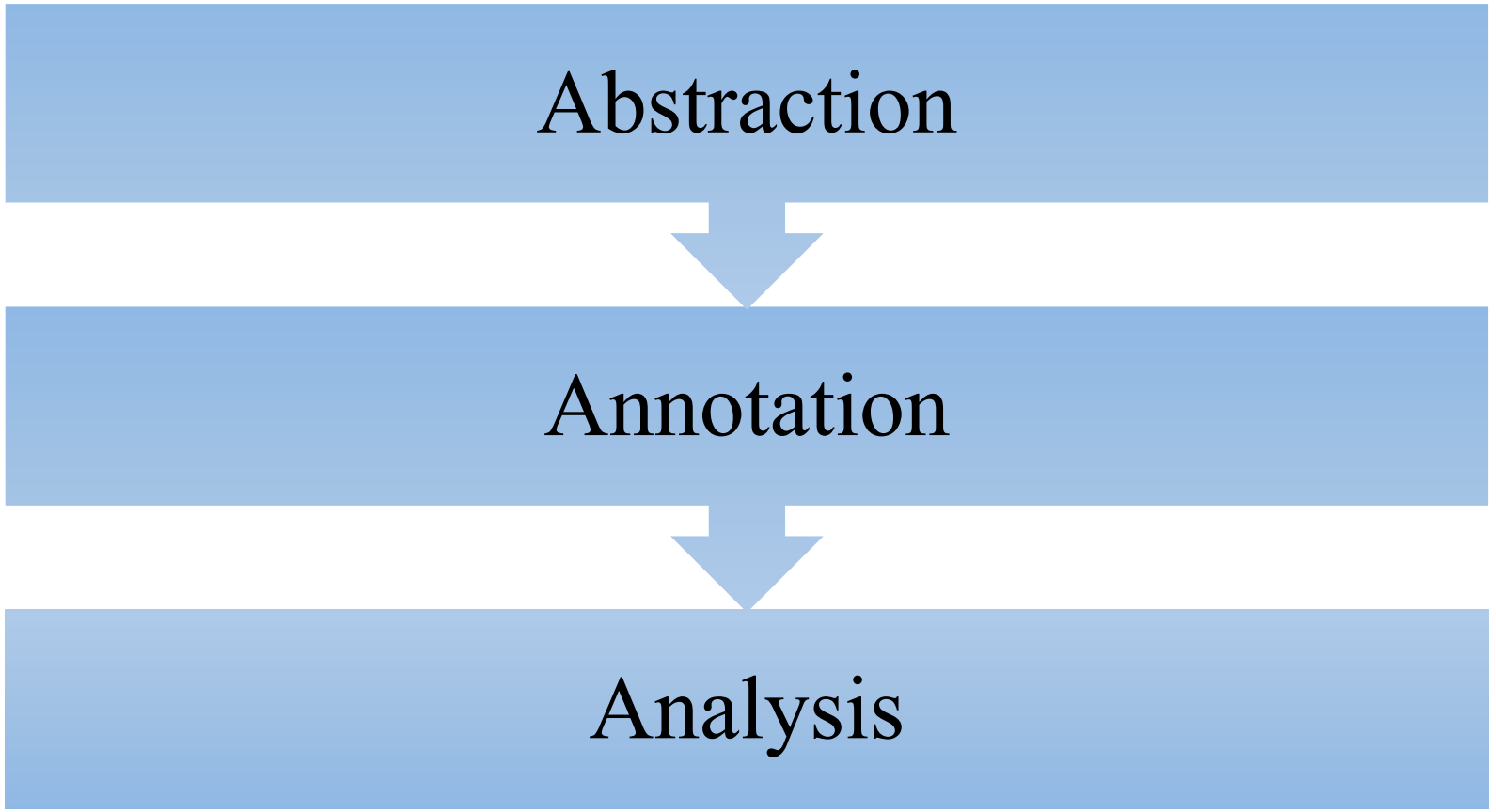
Cumulative connected car innovation activity by auto makers, 2009–15



Source: Das Auto Institut Index, www.auto-institut.de

© 2015 PwC. All rights reserved.

Possibilities of using a corpus linguistic perspectives (three-A models) in compiling a dictionary for specific purposes



Methodological challenges in building up a corpus

1. How to construct a corpus for special purposes adhering authenticity?
2. How to classify corpuses for special purposes?
3. How to construct a bilingual corpus: English-Mongolian adhering translation equivalence?
4. How to insert the database including source and target language /translated versions in language corpus?

There are approximately 5000 codes used in common out of 20,000 OBD codes. Thus, we attempted to collect the most popular codes based on corpus linguistic perspectives- three A models as in the followings;

Example 1: Хуваарилах голын байрлалын мэдрүүрийн хэлхээний гэмтэл 1

1. P0340 Camshaft Position Sensor Circuit Malfunction (Bank 1)
2. P0341 Camshaft Position Sensor Circuit Range/Performance (Bank 1)
3. P0342 Camshaft Position Sensor A Circuit Low Input (Bank 1)
4. P0343 Camshaft Position Sensor A Circuit High Input (Bank 1)
5. P0344 Camshaft Position Sensor A Circuit Intermittent (Bank 1)
6. P0345 Camshaft Position Sensor A Circuit Malfunction (Bank 2)
7. P0346 Camshaft Position Sensor A Circuit Range/Performance (Bank 2)
8. P0347 Camshaft Position Sensor A Circuit Low Input (Bank 2)
9. P0348 Camshaft Position Sensor A Circuit High Input (Bank 2)
10. P0349 Camshaft Position Sensor A Circuit Intermittent (Bank 2)

(Munkhtuya.B, Davaanyam. Sh, Automobile dictionary-2016, page 509)

5 stages for constructing a database for the automobile dictionary

1. Eliminate indexes and index numbers preceding the codes
2. Define the key words in each code
3. Insert the key words as a head word in the database
4. Add compound words and phrases with their definitions
5. Translate the terms and expressions from English into Mongolian language

The importance of extending database

The vital importance of constructing a corpus for specific purposes is availability of language corpuses. A dictionary for specific purposes based on language corpuses does not only have tangible results with its actuality and real life content but also it may help us teach and learn new languages.

Therefore, we can use internet corpus, available language corpus and any open sources.

Megyeri, Marta. Corpus Lexicography.2014.

URL:http:/hdl.handle.net/2437/192184.

Internet corpus (http://www.motorera.com/dictionary)

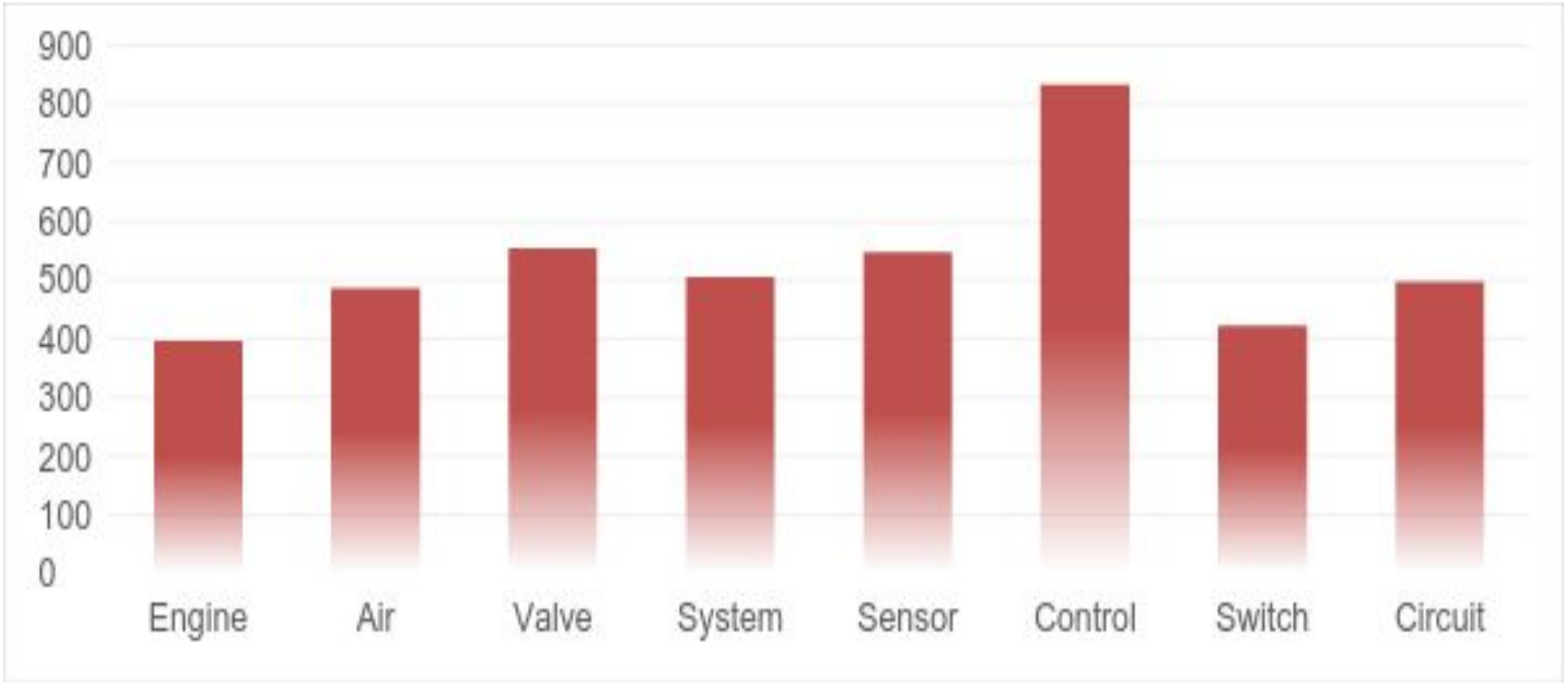
Number of technical terms and terminologies in automobile		Number of technical terms and terminologies in automobile	
A	1598	N	322
B	1940	O	486
C	2252	P	1336
D	1044	Q	255
E	787	R	1008
F	1034	S	2241
G	483	T	1277
H	806	U	179
I	672	V	501
J	111	W	514
K	134	X	13
L	719	Y	24
M	793	Z	41
Total number		**20,675**	

British National Corpus (BNCWeb)

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written. The latest edition is the *BNC XML Edition*, released in 2007.

1st edition, 1994
2nd edition,2001
3rd edition, 2007

Picture 1 : Automobile database: Frequency number of terms and expressions



”**Circuit”** буюу “**хэлхээ**”гэсэн үг нь нэг утгатай нийт 2627 удаа

Зураг 1: ***Circuit /Хэлхээ/ үгийн давтамж***

Frequency breakdown of word and tag combinations for position "node" (1 type and 2627 tokens)

|< << >> >| Frequency breakdown of tags only Go! Download whole table

No.	Word and POS-tags	No. of occurrences	Percent
1	circuit NN1	2627	100%

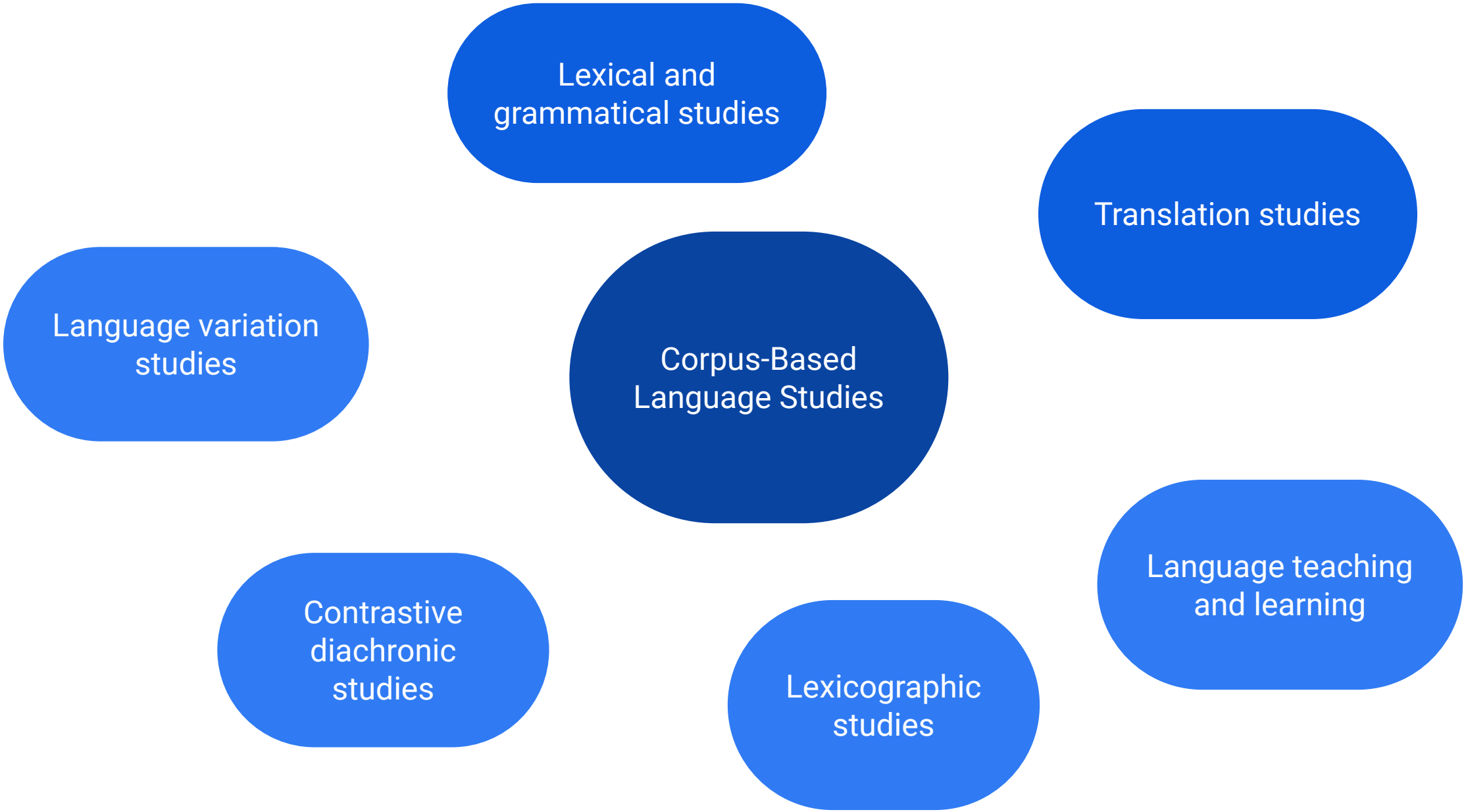
Conclusion

- To sum up, we successfully constructed a corpus for automobile sector with 22.322 professional terms and expressions, 2.264 abbreviations, as well as 2.124 OBD codes in total of 26.710 words.
- Obviously, we need to meet some pre-condition requirements regarding the research work findings as in the followings;

1. Choose appropriate approaches and software programs
2. Apply modern trends and global knowledge
3. Select latest authentic materials which are widely used adhering international standards in the field
4. Collaborate with experts and have professionals do editing and translation, respectively.

Benefits of Corpus-Based Language Studies

- Ease with word processing
- quick to update database
- more reliable without spelling mistakes
- more authentic


Lexical and grammatical studies
Translation studies
Language variation studies
Corpus-Based Language Studies
Language teaching and learning
Contrastive diachronic studies
Lexicographic studies

References

Aarts, B.2001. Corpus linguistics, Chomsky and fuzzy tree fragments in C. Mair and M. Hundt (eds) Corpus Linguistics and Linguistic Theory, pp.5-13.Amsterdam: Rodopi.

Biber, D.1993., Representativeness in corpus design. Literary and Computing 8/4: 243-257.

Carter, R and McCarthy, M. 1997. Exploring Spoken English .Cambridge: Cambridge University Press.

Kucera, H. and Francis, W. 1967.Computational Analysis of Present-day English. Providence: Brown University Press.

Leech, G. 1997. Introducing corpus annotation in R. Garside, G.Leech and A. McEnery (eds) Corpus Annotation, pp.1-18.London: Longman.

Navigli, Roberto. 2009. Using Cycles and Quasi-Cycles to Disambiguate Dictionary Glosses. Proc. of 12th Conference of the European Association for Computational Linguistics (EACL 2009), Athens, Greece, pp. 594-602

Nielsen, Sandro (2008), "The Effect of Lexicographical Information Costs on Dictionary Making and Use", *Lexikos*, 18: 170–189

Nielsen, Sandro (2010): Specialised Translation Dictionaries for Learners. In: P. A. Fuertes-Olivera (ed.): *Specialised Dictionaries for Learners*. Berlin/New York: de Gruyter, 69-82

Tony McEnery; Richard Xiao and Yukio Tono, Corpus-Based Language Studies: Corpus Linguistics: the basics

p. 03-12, p. 208-225, 2008


Thank
you!