Professional and Academic Discourse: an Interdisciplinary Perspective

AN APPRAISAL OF ENRIQUE ALCARAZ'S DESCRIPTION OF ESP LEXICAL TRAITS

María José Marín Pérez & Camino Rea Rizzo

Universidad de Murcia & Universidad Politécnica de Cartagena

34th International Conference of the Spanish Association for Applied Linguistics

14-16th April, Alicante

Contents

Introduction

- **Diverse general definitions of specialised languages**
  - Biber (1988), Halliday (1988): Functional registers displaying specific recurrent features.
  - Cabré (2003): Subcodes of general language determined by pragmatic variables.
  - Sager *et al*. (1980), Alcaraz (2000): Specific language systems used as vehicles of communication amongst specialists in a given field.

- **Enrique Alcaraz's (2000) *El inglés profesional y académico***
  - Fundamental and comprehensive work in the area.
  - Definition of the most relevant traits of EPAP (English for Professional and Academic Purposes):
    - Lexicon, syntax, semantics, pragmatics.

Introduction

Main aims of this research

- A tribute to Alcaraz's seminal work in the area applying a corpus-based methodology: fully automatic method.
- *TC*: 1.2 millon-word Telematics English corpus (Rea, 2008).
- *UKSCC:* 2.6 million-word legal English corpus (Marín, 2014).

Major research foci

- Automatic extraction and comparison of the terminology in both corpora (ratio and distribution).
- Relevance of sub-technical terms in both varieties.
- Use of Latin words and phrases in *UKSCC*, the legal corpus.
- Incidence of acronyms and abbreviations in *TC*, the telematics corpus.

Specialised Terminology in *TC* and *UKSCC*

- One of the most relevant features of EPAP is the use of specialised terminology:
  - Cabré (2000: 62): «form and content units which, used in different discursive conditions, acquire a specialised value»
  - Alcaraz (2000): Univocality of terms. Key to the understanding of specialised texts.

- Large text collections: Automatic Term Recognition (ATR) methods. Some of them were implemented on *TC* and *UKSCC* to assess their precision levels:
  - *TF-IDF* (Sparck Jones, 1972).
  - *TermoStat* (Drouin, 2003).
  - *C-Value* (Frantzi and Anniadou, 1999).
  - *Terminus* (Nazar and Cabré, 2012).

Specialised Terminology in *TC* and *UKSCC*

- Results of ATR method assessment

PRECISION %

***Terminus* (Nazar & Cabré, 2012)**

UKSCC	***TC***
84.5% (top 200)	69.5% (top 200)
71.5% (average)	60% (average)

1,787 terms extracted from *UKSCC*

888 extracted from *TC*

- Term frequency and distribution

UKSCC (193 texts)

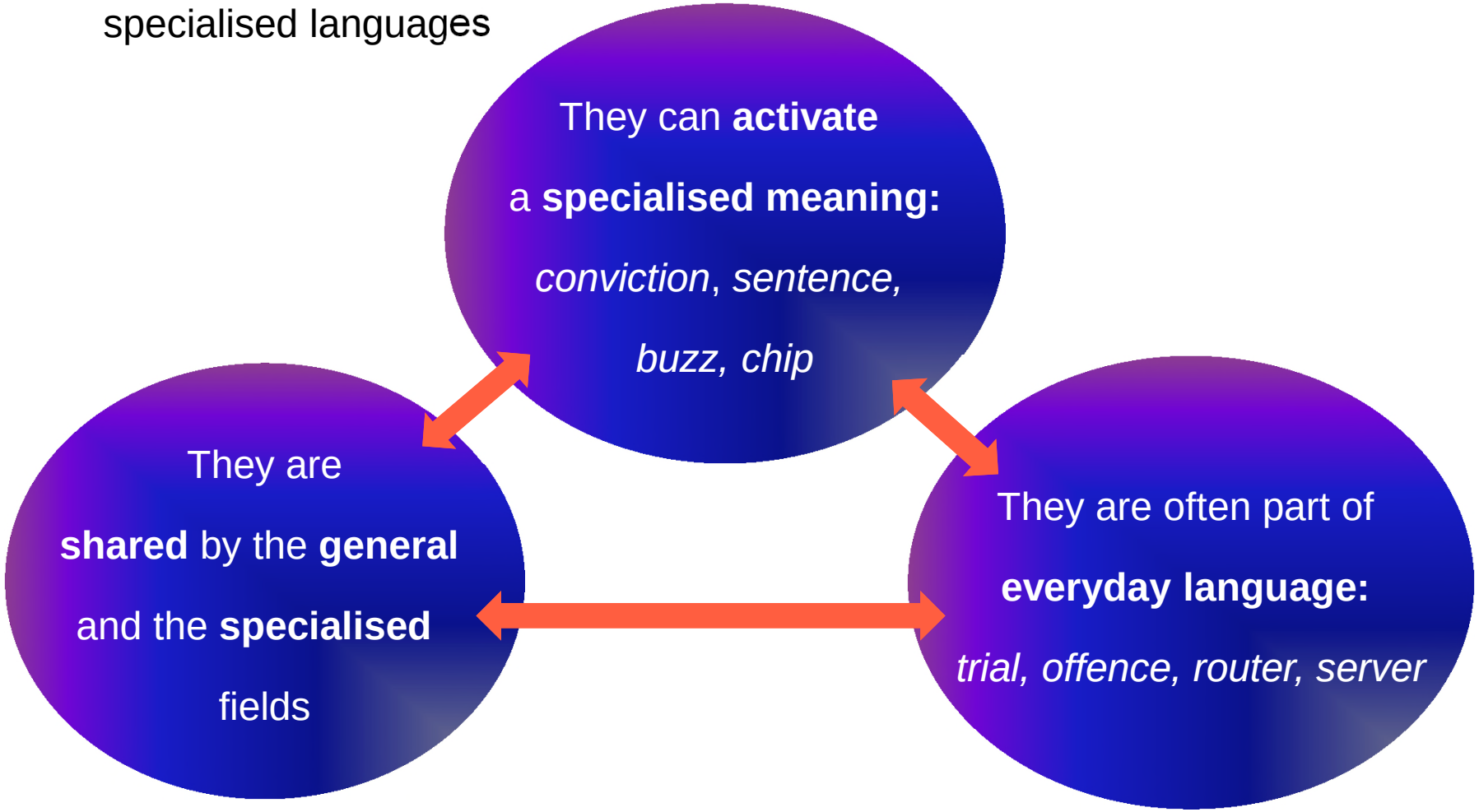
TC (272 texts)

	Av. Freq.	**Distribution**
***UKSCC* terms**	1,037	14%
***UKSCC* types**	169.45	10%
***TC* terms**	38.62	11%
TC types	89.93	5.36%

Alcaraz's observation as regards the **relevance** and **significance** of specialised terminology is thus **supported** by **stastistical data**.

Subtechnical terms in *TC* and *UKSCC*

- Alcaraz (2000) also highlights the **relevance** of **subtechnical** vocabulary in specialised languages



Subtechnical terms in *TC* and *UKSCC*

TC and ***UKSCC*** term inventories against ***BNC*** most frequent 3,000 English words

	Overlapping terms	Percentage
UKSCC	810/1787	47.35%
TC	315/888	35.55%

- **High incidence** of **overlapping** terms between both lists.
- Terms like *action*, *claim*, *criminal* or *processor* and *controller* were found amongst them.
- These **data prove** Alcaraz's statement on the **significance** of **subtechnical** terms both legal and telecommunications English.

Latin terms in Legal English

- **Involved parameters**
  1. Frequency
  2. Text range
  3. Keyness
  4. Degree of specialization

- Focus on **purely Latin borrowings** like *obiter dictum* or *ratio decidendi*, which were imported directly from Latin without being adapted into English.
- A list of Latin terms from text and academic books acted as reference for the identification of these lexical units in *UKSCC*.

Latin terms in Legal English

- **Involved parameters**

1. Frequency: number of times a word is count in the corpus.
   - **187 single word Latin units** extracted.
     - Top 10 most frequent ones were function words i.e.: *versus (v), per, de, inter, re.*
     - Low frequency: 400th and 1800th positions of the frequency rank.
     - Only 17 terms found within the top 2000 word types i.e.: *affidavit, quantum, jure* or *incapax*.
2. Text range: the percentage of running words in a text covered by a term.
   - 187 Latin terms in comparison to 35 crime nouns (*murder, abduction, threats, battery*, etc).

	Frequency	Text range
Latin terms	Low	0.0059%
Crime nouns	Higher	0.00095%

Latin terms in Legal English

- **Involved parameters**

  3. Keyness (given by *Wordsmith*): "A word is considered key if it is unusually frequent (or infrequent) in comparison with what one would expect on the basis of the larger word-list" (Scott, 2008).

	Average keyness
Latin terms	94.3
Crime nouns	97.3
Whole list	116.08

Latin terms in Legal English

Involved parameters

4. Degree of specialisation: Application of Chung's method (2003).

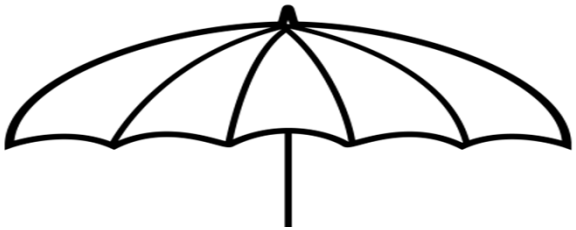
- A word type is classified as a term only "if it occurs 50 times more often in the technical text than in the comparison corpus, or if it only occurs in the comparison corpus" (2003: 53).
- Terms whose ratio >50: *affidavit, caveat, proviso, extempore, quantum, lex* or *subpoena.*
- Most terms are part of academic or general vocabulary (*plus, nil, persona, memorandum, caveat* or *alibi)* or they simply do not occur in isolation but rather as part of phrases → ***Latin phrases***.

Ratio = ∞	**38%**	*Ex turpi causa, Doli incapax, Quantum meruit, Pari delicto, Res iudicata*
Ratio >50	**20%**	*De jure, A fortiori, Ultra vires, Ex parte*
Ratio <50	**21%**	*De facto, Ipso facto, Sui generis, Vice versa, per se*

Abbreviations in Telecom. English

- **Shortening processes**
  - New lexical items may be formed by ***deleting*** linguistic material: **abbreviation** of existing words and expressions, omission of elements in compounds, creation of letter symbols, combination of letters and numbers into short designations, and pictograms (Sager et al., 1980).

Abbreviation is an umbrella term referring to any kind of word which has undertaken a shortening process, that is, any compressed form in general.



Initialism or initials:

TNT, DVD, GPS.

Acronym:

NASA, NATO; LASER.

Clipping:

doc from *doctor*, *sec* from *second*, *gbyte* from *gigabyte.*

Abbreviations in Telecom. English

- **Overall quantitative behaviour**
  - **Telecommunication Engineering Word List (TEWL)** (Rea, 2008).
    - **Abbreviations** stand for **16%** of TEWL.
    - TEWL includes the most salient, central and typical specialized lexical units in the domain.
    - Both words whose use is restricted to the subject domain and those which activate a specialized meaning in the discipline.
    - **2,747 specialized lexical units** (402 whole technical families) within the range of the 1000 most statistically significant word families in the domain.
    - Comparative approach: General language corpus **LACELL** (20 millions) and the corpus specialized in Telecommunication Engineering English (**TEC**) (5.5 million words).

Abbreviations in Telecom. English

- **Overall quantitative behaviour**
  - **Abbreviations in TEC,** the main Telecommunications Corpus.
    - **443** abbreviations comprise **16%** of the terms in TEWL.
    - Text range: 15% of the tokens covered by TEWL.
  - **Abbreviations in TC,** the subcorpus of Telematics.
    - **32** abbreviations recognised as terms by *Terminus* ATR method.
    - Text range: 2.7% of the tokens covered by TEWL.

Abbreviations in Telecom. English

Relevance of compressed forms

- Sample of abbreviations in TEWL and quantitative parameters.

Rank	TEWL	F Tec	F Lacell	Ratio	Term	Keyness
1	IP	5,239	20	994,85	spc	16,182
2	**TCP**	1,717	12	543,41	spc	5,248
3	**ATM**	1,639	35	177,85	spc	4,817
4	LAN	1,481	27	208,32	spc	4,387
5	OSPF	1,284	0	∞	inf/spc	4,027
6	QOS	1,155	0	∞	inf/spc	3,622
7	VHDL	1,150	0	∞	inf/spc	3,607
8	**MPLS**	1,112	0	∞	inf/spc	3,487
9	GSM	1,109	4	1052,96	spc	3,427
10	**VPN**	1,007	5	764,89	spc	3,097
11	**IEEE**	1,002	9	422,83	spc	3,044
12	**LSAS**	858	1	3258,58	spc	2,676
13	DSP	906	41	83,92	spc	2,523
14	**LSA**	804	0	∞	inf/spc	2,521
15	CDMA	805	1	3057,29	spc	2,510
16	CISCO	840	14	227,87	spc	2,498
17	**MHZ**	792	18	167,11	spc	2,319
18	GHZ	734	2	1393,82	spc	2,275
19	FPGA	713	0	∞	inf/spc	2,236
20	SCTP	703	0	∞	inf/spc	2,205
21	RF	716	8	339,91	spc	2,161
22	DB	774	36	81,65	spc	2,149
23	WLAN	677	0	∞	inf/spc	2,123
24	ISDN	699	14	189,62	spc	2,061
25	**HTTP**	801	96	31,69	no	1,946

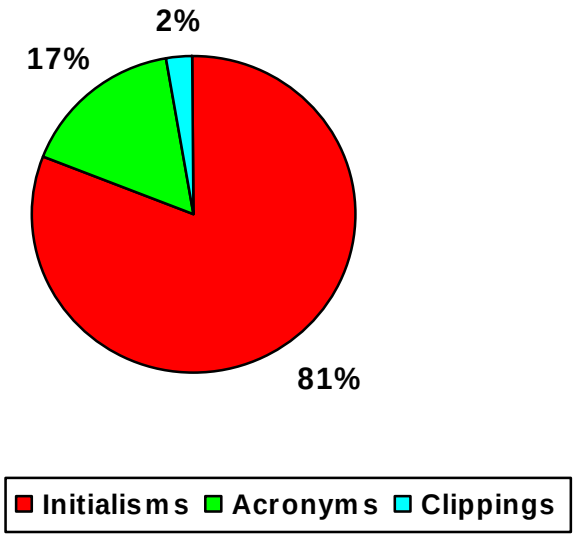
Abbreviations in Telecom. English

- **Relevance of compressed forms**
  - Keyness and specialization of the 443 abbreviations in TEWL:

Abbreviation	Ratio	Term	Keyness
237 = 53%	∞	Not found in general language	***OSPF***: 4,027 ***VDMS***: 12.5
119 = 27%	>50	High freq. in TEC and low freq. in LACELL	***IP***: 16,182 ***HDLS***: 63
87 = 20%	<50	Use extended to general language	***HTTP***: 1,946 ***GUIS***: 11

Abbreviations in Telecommunications

- **Classification depending on the compression process**
  - **Initials** (360): *IP, TCP, ATM, GPRS, SNMP, BGP, DCE, GPS, IGRP, PBX, BS, RMI, DHCP, LSP, LSPS, UML, ISP, ICT, IPX, PDU, CPE, DD, MSC,* etc.
  - **Acronyms** (74): *RADARS, VOIP, PAC, QOS, MAC, OSI, LABVIEW, SPICE, SONET, CORBA, WAP, ASIC, GIS, SIM, LEDS, MIMO, ANSI, MOSFET,* etc.
  - **Clippings** (11): *LOG* (logarithm), *CONFIG* (configuration), *SYNC* (synchronization), *GBYTE* (gigabyte), *DEMUX* (demultiplexer), *MUXS* (multiplexers), *GAAS* (gallium arsenide).



Abbreviations in Telecommunications

- **Lexicalised abbreviations**
  - Abbreviations as full words capable of compounding, derivation and conversion.
    - Abbreviations in singular and plural: *LAN/S, VLAN/S, RAM/S, COMSAT/S, RADAR/S, FIFO/S, PAC/S,* etc.
    - Compounding, multiword units and shortening again: **metal-oxide semiconductor (MOS)**
      - *CMOS* (complementary metal-oxide semiconductor)
      - *NMOS* (n-channel metal-oxide semiconductor)
      - PMOS (p-channel metal-oxide semiconductor)
      - *BICMOS* (bipolar complementary metal-oxide semiconductor)
      - *MOSFETs* (metal-oxide semiconductor field-effect transistors)
    - Abbreviations integrate longer word combinations: ***QoS* (quality of service)**
      - QoS architecture / end-to-end QoS / QoS capabilities
      - dynamic QoS parameters / on-demand QoS routing protocol / inter-domain QoS signalling protocol / adaptive QoS control scheme

Final remarks

- The adoption of a corpus-driven approach has allowed to identify automatically the typical behaviour of the lexical characteristics which define special languages.

- The application of ATR methods and quantitative parameters have corroborated Alcaraz's description of a set of lexical features in ESP:
  - The use of specialised terminology
  - The relevance of subtechnical terms
  - The particular use of Latin terms and phrases in legal English
  - The outstanding presence of abbreviations in Telecommunications English

Professional and Academic Discourse: an Interdisciplinary Perspective

PRAISING ENRIQUE ALCARAZ'S DESCRIPTION OF ESP LEXICAL TRAITS

María José Marín Pérez & Camino Rea Rizzo

Universidad de Murcia & Universidad Politécnica de Cartagena

34th International Conference of the Spanish Association for Applied Linguistics

14-16th April, Alicante

References

- Alcaraz Varó, E. (2000). *El inglés profesional y académico*. Madrid: Alianza Editorial.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Cabré, M.T. (2000). "Terminologie et linguistique: la théorie des portes". *Terminologies nouvelles. Terminologie et diversité culturelle,* 21: 10-15.
- Cabré, T. M. (2003). Terminology. Theory, methods and applications. Terminology 9 (2): 163–199.
- Drouin, P. (2003). "Term extraction using non-technical corpora as a point of leverage". *Terminology* 9, 1: 99-117.
- Frantzi, K.T. and Ananiadou, S. (1999). "The c/nc value domain independent method formulti-word term extraction". *Journal of Natural Language Processing* 3 (2): 115-127.
- Halliday, M. (1988). "On the language of physical science". In Ghadessy (Ed.): 162-167.
- Marín, M.J. (2014). "Evaluation of five single-word term recognition methods on a legal corpus". *Corpora,* 9 (1). Endinburgh: Endinburgh University Press.
- Nazar, R., Cabré, M.T. (2012). "Supervised Learning Algorithms Applied to Terminology Extraction" in Aguado de Cea, G., Suárez-Figueroa, M.C., García-Castro, R., Montiel-Ponsoda, E. (eds.), *Proceedings of the 10th Terminology and Knowledge Engineering Conference* (TKE 2012). Madrid: Ontology Engineering Group, Association for Terminology and Knowledge Transfer.
- Rea, C. (2008). *El inglés de las telecomunicaciones: estudio léxico basado en un corpus específico*. Tesis doctoral. Murcia: Servicio de Publicaciones de la Universidad de Murcia.
- Sager, J.C., Dungworth, D., McDonald, P.F. (1980). *English Special languages*. *Principles and practice in science and technology.* Wiesbaden, Brandstetter Verlag KG.
- Sparck Jones, K. (1972). "A statistical interpretation of term specificity and its application in retrieval". *Journal of Documentation* 28: 11-21.