Facing the Identification Problem in Language-Related Scientific Data Analysis.

Joseph Mariani[1,2], Christopher Cieri[3], Gil Francopoulo[1,4],
Patrick Paroubek[2], Marine Delaborde[2]
[1]IMMI, [2]LIMSI, [3]LDC, [4]Tagmatica
[1,2]BP 133 - 91403 ORSAY cedex (France)
[3]Linguistic Data Consortium, 3600 Market Street, Philadelphia, PA 19104 (USA)
E-mail: joseph.mariani@limsi.fr, ccieri@ldc.upenn.edu, gil.francopoulo@wanadoo.fr, pap@limsi.fr,
mdelaborde@yahoo.fr

Abstract

This paper describes the problems that must be addressed when studying large amounts of data over time which require entity normalization applied not to the usual genres of news or political speech, but to the genre of academic discourse about language resources, technologies and sciences. It reports on the normalization processes that had to be applied to produce data usable for computing statistics in three past studies on the *LRE Map*, the *ISCA Archive* and the *LDC Bibliography*. It shows the need for human expertise during normalization and the necessity to adapt the work to the study objectives. It investigates possible improvements for reducing the workload necessary to produce comparable results. Through this paper, we show the necessity to define and agree on international persistent and unique identifiers.

Keywords: Data identification, Normalization, ISLRN, Archive analysis, Bibliometrics, Scientometrics

1. Introduction

Several initiatives took place recently concerning the analysis of data in the literature related to research in Natural Language Processing (NLP) (ACL, 2012), Spoken Language Processing (SLP) or Language Resources (LR). The authors have conducted such studies on the FLaReNet *LRE Map* (N. Calzolari et al., 2012), on the *ISCA Archive* (J. Mariani et al., 2013) and on technical papers mentioning the LRs distributed through LDC (E. Ahtaridis et al., 2012). All have experienced difficulties in correctly identifying the referents mentioned in the data they wished to study and invested significant manual effort to permit proper analysis. Difficulties concerned many aspects: authors' name, affiliation, gender and nationality, the citation title, source and acknowledgments, the technical terms in the papers and the Language Resources' name and type, the languages they cover, the modalities they address, etc.

Our experiences relate to the general problem of reference resolution in the context of scientific literature information mining, named entity annotation, or on a more general basis, the situation where a datum must be normalized to designate a unique referent.

2. Problem Classification

Four kinds of problem arise depending on the data cleaning performed:

1. The first concerns **identification in synchrony**, in other terms, to identify a specific referent at a given point in time. There are two sub-cases: i) selecting a standardized name in a predefined list, for instance the name of a country, ii) standardizing a name when such list does not (currently) exist, for instance when dealing with authors or titles of scientific papers. For the first situation, the obvious and best strategy is to use standardized codes such as ISO-3166-1 alpha-3. Defining ones own list of codes is not a good option for many reasons but also because difficult problems arise for example around national status: are "Palestine", "Kurdistan", "Basque country" or a group like the "Baltic countries" to be considered? Such difficult problems are beyond the present study. If a non-standardized option is adopted, explaining the rationale is lengthy and complex. The problem of country name identification is just one example: ISO and W3C maintain lists of codes for various domains like language identification (see ISO-639-3), linguistic properties like grammatical gender (see ISO-12620)[1].

Concerning person names, it should be noted that, at the moment, there is no standardized list, though there are multiple proposals, and normalizing these names is a serious challenge. Different people have the same name and although the number of homographs is rather small within a specific scientific domain, the boundaries of scientific domains are not sharp, nor are they recognized by current search tools. Some cleaning must be done when comparing names of different length, like "Robert Smith" compared to "Robert D Smith" or variations with titles like "Sen. Robert Smith" or "Robert Smith Jr".

[1] with the register located at www.isocat.org.

2. The second problem concerns **name changes over time (diachrony)**. Different sub-cases arise like a simple name change (e.g. “Belgian Congo” becoming “Democratic Republic of the Congo”), a division (e.g. “Czechoslovakia” split into “Czech Republic” and “Slovak Republic” or “Slovakia”), a merger (e.g. “German Democratic Republic”, informally “East Germany” and “Federal Republic of Germany”, informally “West Germany” merged into “Federal Republic of Germany”, informally now “Germany”), or possibly some slightly more complex situations which appear when cities are reorganized (e.g. in Quebec, the former city “Hull” has been merged with the suburban city “Gatineau” and the new association is called “Gatineau”). These phenomena should be handled in a coherent manner to allow consistent counting and annotation over the period. In some domains, ISO maintains a list specifically for this purpose, (e.g. ISO-3166-3 that defines codes and rules for countries deleted from ISO-3166-1 since its first publication in 1974). A person may change names over time too, for example at marriage in many cultures.

3. The third problem concerns **reference precision**. The challenge is to name the intended item, group or abstraction, for instance, “WordNet” compared to “WordNet 1.5” or “WordNet 3.0”. The criteria for selecting the right level of designation reflect the precision needed for the study, the information available, and of course, the precision coherence from one resource to another.

4. The forth problem is of a different nature and is more practical. It concerns the **mismatch between the appearance within input documents and the semantics of the references they contain**. For instance, when the entry is a PDF file without the required fonts, the word segmentation is frequently erroneous. Another example is hyphenation (i.e. sentence splitting at the end of a line), which produces some small noisy strings difficult to fix automatically. Handling these problems is often difficult, and a hybrid strategy combining generic mechanisms like rules or data-driven behavior with specific exceptions is often needed.

3. Our experience in data normalization within analytic studies

We will now provide some examples of the problems the authors faced and the solutions chosen in three different studies: those of the *LRE Map*, of the *ISCA Archive* and of the *LDC Bibliography*.

3.1. The *LRE Map*

The *LRE Map* is produced by the authors of papers submitted to various conferences, starting with LREC 2010. Authors provide information on the language resources mentioned in their paper, whether produced or used, when they submit it (N. Calzolari et al., 2012). The *LRE Map* yielded the *Language Matrices* within the T4ME project (N. Calzolari and J. Mariani, 2012). They provide the number of Language Resources by type and modality (text, speech, multimodal / multimedia) in various languages, including *Sign Languages*. They highlight the lack of resources for certain languages and thus can serve to guide efforts to equalize and complete coverage. They may also help computing a Language Resource Impact Factor (LRIF) in order to recognize the contribution to research of LR producers just as publications impact factors recognize the influence of papers authors.

The parameters of the study are the resource Type, Name, Modality and Language coverage.

3.1.1. *LRE Map* Processing

The analysis in our first study of the data gathered at LREC 2010 revealed that the same Language Resources were mentioned and described differently by different authors requiring human experts to normalize and map the information. The acquisition software was then improved and used to complete the initial *LRE Map* with the data coming from nine more conferences on Speech and Natural Language Processing until LREC 2012.

Just as for the initial *LRE Map* in 2010, it was also necessary to clean up the resulting data in order to produce a new set of *Language Matrices*. Even if the process was less demanding this time, it was nonetheless very tedious and ostensibly unnecessary.

The data is kept as entered by the authors, including any errors they may have introduced, and the clean version of the information is added in parallel. The processed information is used for all analyses, but the raw data is provided to the users who wish to consult the database, leaving the full responsibility of the information to the authors.

Cleaning and modifying the data as a collaborative activity is very difficult, as some decisions are specific to the intended uses. The modifications are related to various issues (cases, spelling errors, introduction of a new term while the concept was already covered by or within an already existing term, etc.), but also raise in some cases more fundamental and theoretical questions, which should be discussed within the scientific community.

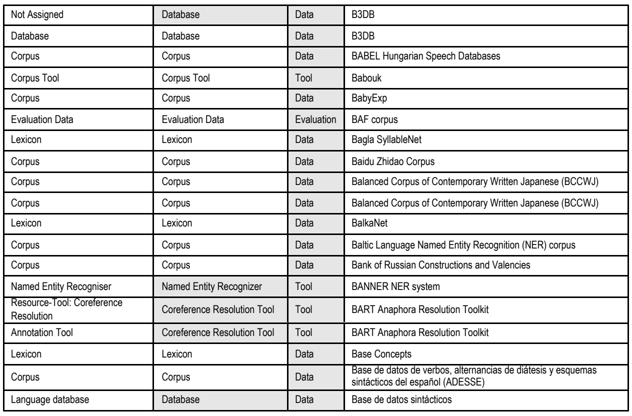
The data which were corrected are the following:

3.1.1.1. Resource Type

We considered 4 categories of LR Types: Data, Tools, Evaluation and Meta-Resources (metadata, standards, guidelines, etc.), and various Types within those categories (such as corpus, lexicon, parser, evaluation package, or standards) that we suggested to the authors, allowing them to enter other Types than those we suggest.

The errors here concern for example new Types that were introduced by the authors though appropriate Types already existed but were perhaps missed by the authors.

The study of the new LR Types provided by the authors also reveals Types that were not anticipated by the *LREC Map* designers, due to the difficulty of embracing the whole field. It may then happen that different authors use different terms for the same new Type, that should therefore be harmonized, or that the Type is too specific and can be merged in a larger Type (already suggested or new). They also allow one to identify new kinds of research conducted somewhere in the world on a given language, and act as a “weak signal” indicating new emerging research areas in the field of language science and technology.



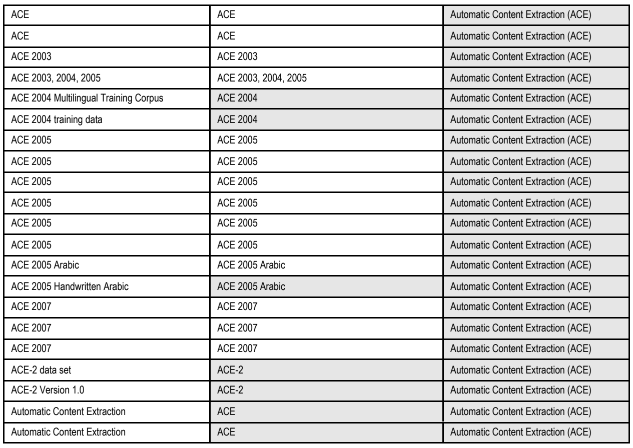
Not Assigned	Database	Data	B3DB
Database	Database	Data	B3DB
Corpus	Corpus	Data	BABEL Hungarian Speech Databases
Corpus Tool	Corpus Tool	Tool	Babouk
Corpus	Corpus	Data	BabyExp
Evaluation Data	Evaluation Data	Evaluation	BAF corpus
Lexicon	Lexicon	Data	Bagla SyllableNet
Corpus	Corpus	Data	Baidu Zhidao Corpus
Corpus	Corpus	Data	Balanced Corpus of Contemporary Written Japanese (BCCWJ)
Corpus	Corpus	Data	Balanced Corpus of Contemporary Written Japanese (BCCWJ)
Lexicon	Lexicon	Data	BalkaNet
Corpus	Corpus	Data	Baltic Language Named Entity Recognition (NER) corpus
Corpus	Corpus	Data	Bank of Russian Constructions and Valencies
Named Entity Recogniser	Named Entity Recognizer	Tool	BANNER NER system
Resource-Tool: Coreference Resolution	Coreference Resolution Tool	Tool	BART Anaphora Resolution Toolkit
Annotation Tool	Coreference Resolution Tool	Tool	BART Anaphora Resolution Toolkit
Lexicon	Lexicon	Data	Base Concepts
Corpus	Corpus	Data	Base de datos de verbos, alternancias de diátesis y esquemas sintácticos del español (ADESSE)
Language database	Database	Data	Base de datos sintácticos

Fig. 1. Examples of cleaning the inputs of the LR Types (first column): Spelling normalization (“recognizer”), Harmonization (BART as a “Coreference Resolution Tool”), “Language Database” as “Database” (second column) and categorization into “Data”, “Tool”, “Evaluation” or “Meta-Resources” gross categories (third column). The entered LR names appear in the fourth column.

In the initial 2010 *LRE Map*, we suggested 24 widely recognized Types that covered 85% of the entries. However, the remaining 15% represented a long tail of 222 “Other” Types, including a small set (13) of resources belonging to several Types (1%), 100 Other Types mentioned several times by authors (7%) and 99 Types mentioned only once (7%). The introduction of 5 new Types and of the auto-completion process improved the homogeneity of the data and reduced the number of new Types introduced by the authors. In the final complete *LRE Map*, the suggested Types cover 90% of the entries. Only 149 Types are not among the suggested ones, the most frequent ones being *Machine Learning Tool* (mentioned 21 times), *Database* (16), *Terminology Tool* (10) and *Search Engine* (9).

3.1.1.2. Resource Name

In the initial *LRE Map*, the LR name was provided by the authors in an open format. Therefore different authors may use different wordings for the same LR: it may be mentioned by its full name, abbreviation or acronym. The mention may contain the date or the version number. The problem of versioning is a difficult problem to handle: should the different versions of a LR over time be considered as the same LR or as different ones? Some surveys may wish not to distinguish between the various versions of a Language Resource. However, because even the versions may be labeled inconsistently, combining them is non-trivial. For the time being, we decided to keep them as different items. Also a LR may be constituted by several parts (corresponding to different Types: a dictionary and a corpus, for example) under a single name or have varieties in different languages (Wordnets for example). In some cases, we kept them separate. Furthermore, once two language resources have been declared identical, it is necessary to check if their types, modalities and languages are the same and resolve any difference. Overall our goal was to normalize the various renderings of an LR name into a single handle that could then be related to the official name of the resource where present.



ACE	ACE	Automatic Content Extraction (ACE)
ACE	ACE	Automatic Content Extraction (ACE)
ACE 2003	ACE 2003	Automatic Content Extraction (ACE)
ACE 2003, 2004, 2005	ACE 2003, 2004, 2005	Automatic Content Extraction (ACE)
ACE 2004 Multilingual Training Corpus	ACE 2004	Automatic Content Extraction (ACE)
ACE 2004 training data	ACE 2004	Automatic Content Extraction (ACE)
ACE 2005	ACE 2005	Automatic Content Extraction (ACE)
ACE 2005	ACE 2005	Automatic Content Extraction (ACE)
ACE 2005	ACE 2005	Automatic Content Extraction (ACE)
ACE 2005	ACE 2005	Automatic Content Extraction (ACE)
ACE 2005	ACE 2005	Automatic Content Extraction (ACE)
ACE 2005	ACE 2005	Automatic Content Extraction (ACE)
ACE 2005 Arabic	ACE 2005 Arabic	Automatic Content Extraction (ACE)
ACE 2005 Handwritten Arabic	ACE 2005 Arabic	Automatic Content Extraction (ACE)
ACE 2007	ACE 2007	Automatic Content Extraction (ACE)
ACE 2007	ACE 2007	Automatic Content Extraction (ACE)
ACE 2007	ACE 2007	Automatic Content Extraction (ACE)
ACE-2 data set	ACE-2	Automatic Content Extraction (ACE)
ACE-2 Version 1.0	ACE-2	Automatic Content Extraction (ACE)
Automatic Content Extraction	ACE	Automatic Content Extraction (ACE)
Automatic Content Extraction	ACE	Automatic Content Extraction (ACE)

Fig. 2. Examples of cleaning the inputs of the LR Names (first column): Harmonization of various ways of mentioning the “Automatic Content Extraction (ACE)” resource (second column), and gathering into the same family resource name (third column).

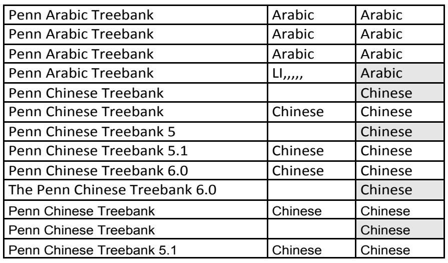
To harmonize the way authors enter resource names, the second version of the *LRE Map* software included auto-completion: the system proposes to the author a list of LR names, based on the characters he/she types. This improvement greatly reduced the workload in the second cleaning operation, and only 8% of the names had to be corrected compared with 45% at LREC 2010. Unfortunately, 64 entries lacked any LR name at all. After normalization, this phase resulted in an updated *LRE Map* comprising 4,295 entries.

Beyond this naming issue, the more general concern is to identify the same LR, and to merge the accidentally varying entries in order to avoid counting the same LR twice. Regarding LREC 2010, after doing this merging, the number of entries only decreased from 1,995 to 1,576 different LRs (about 20%), corresponding to an average ambiguity factor of 1.33 and a standard deviation of 1.25, which shows that the diversity of the LR used was large in this first analysis. For the entire *LRE Map*, it decreased from 4,395 to 3,121 (about 30%), as the production of new LRs may not vary as fast as their use.

This demonstrates the importance of attributing and requiring a Persistent and Unique LR Identifier (LRID). Such an ID would also allow one to trace the use of the corresponding LR in research through publications, and more generally to trace the use, enrichment, or modification of such LRs by all Language Technology stakeholders (P. Labropoulou, C. Cieri and M. Gavriilidou, 2014). Just as a book is identified by an ISBN number, the same should be done for LR, in a coordinated way, through a single entity attributing such numbers. This appears as a big challenge if we consider its international dimension, and discussions are presently on going at the international level in order to define the way to assign an International Standard Language Resource Number (ISLRN) to each Language Resource (K. Choukri et al., 2012). Interestingly, the Biology community is currently conducting similar reflections on Biological Resources (A. Cambon-Thomsen et al., 2011).

3.1.1.3. Resource Language(s)

Here also, the input format was initially unconstrained. It was therefore necessary to review the data in order to harmonize the various spellings. Some authors provided the ISO code for languages, instead of the complete name. Some decisions were harder: we chose for example to consider British English, American English and English as a single (EU) language.



Penn Arabic Treebank	Arabic	Arabic
Penn Arabic Treebank	Arabic	Arabic
Penn Arabic Treebank	Arabic	Arabic
Penn Arabic Treebank	LI,,,,,	Arabic
Penn Chinese Treebank		Chinese
Penn Chinese Treebank	Chinese	Chinese
Penn Chinese Treebank 5		Chinese
Penn Chinese Treebank 5.1	Chinese	Chinese
Penn Chinese Treebank 6.0	Chinese	Chinese
The Penn Chinese Treebank 6.0		Chinese
Penn Chinese Treebank	Chinese	Chinese
Penn Chinese Treebank		Chinese
Penn Chinese Treebank 5.1	Chinese	Chinese

Fig. 3. Examples of cleaning the inputs of the language(s) addressed by a LR (entered input in first column): Harmonization of the "Penn Arabic Treebank" to cover the Arabic language, and of the "Penn Chinese Treebank" to address Chinese (third column).

The new acquisition software provides a list of languages encoded according to the ISO 639-3 codes based on the *Ethnologue* survey of existing languages (P. Lewis, 2009), which facilitates and normalizes the input of the 5 first languages corresponding to a LR. In case an LR addresses more languages, the others are entered in a free format.

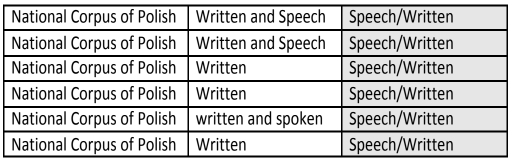
In the first study of the Language Matrices, the focus was on EU languages. We therefore only considered the 23 official EU languages at that time; we merged all the regional EU languages (such as Basque, Catalan, Galician, Sami, Luxemburgish) in a single category ("European Regional languages") and the non-EU European languages (such as Moldavian, Turkish, Norwegian) also in a single category ("Other European languages"). We finally completed the set of categories with "Multilingual", "Language Independent" and "Non Applicable".

In the second study, we considered a set of 4 Spanish regional languages (Catalan, Basque, Galician and Asturian) individually as well as the 23 official EU languages. We gathered all the other European languages in a single "Other European Languages" category (comprising 51 national and regional European languages). We considered individually major international languages (Arabic, Hindi, Chinese Mandarin, Japanese and Korean), and gathered all the other ones in a single "Other Languages" category (comprising 133 languages). This represents a total of 216 languages mentioned in the *LRE Map* entries. We also studied 21 Sign Languages and specifically a set of 27 European regional and minority languages (C. Soria and J. Mariani, 2013).

We decided that LRs which are relevant to several languages should be counted for each of those languages in the analysis of the language coverage.

3.1.1.4. Resource Modality(ies)

Here also, a limited set of suggestions was made to the authors. However, some LRs may be related to multiple modalities (such as spoken and written). In this case, we decided to mark them as such in a coherent way and we considered them for both modalities. Some authors introduced modalities that we considered as already covered (such as "Text" instead of "Written", or "Any modality" instead of "Multimodal/Multimedia"), and we therefore had to correct a few cases.



National Corpus of Polish	Written and Speech	Speech/Written
National Corpus of Polish	Written and Speech	Speech/Written
National Corpus of Polish	Written	Speech/Written
National Corpus of Polish	Written	Speech/Written
National Corpus of Polish	written and spoken	Speech/Written
National Corpus of Polish	Written	Speech/Written

Fig. 4. Examples of cleaning the modality of a LR (second column): harmonization of "National Corpus of Polish" to address both spoken and written language (third column).

3.1.2. Summary of the Normalizations

A final normalization was still required on the full *LRE Map* after merging the data from 10 conferences, in order to ensure the overall homogeneity of Names and Types across conferences. 218 Types and 231 Names (about 5% of the entries) and 6 Modalities were still corrected in this final process.

Conference	Date, place	Entries	Existing Types	Corrected Existing Types	New Types	Corrected New Types	Corrected Names	Corrected languages	Corrected Modalities	# corrections	% corrected cells
LREC 2010	May 17-23 2010, Valetta, Malta	1995	1720 (87%)	156 (8%)	104	15	889 (45%)	179 (9%)	122 (6%)	1361	17%
LREC 2012	May 21-27 2012, Istanbul, Turkey	1130	1012 (90%)	55 (5%)	43	20	95 (8%)	13 (1%)	83 (7%)	266	6%
Total	10 conferen ces	4395	3883 (89%)	283 (6%)	192	38	1149 (26%)	307 (7%)	437 (10%)	2214	13%

Table 1 Table presenting the number of corrections on Types, Names, Languages and Modalities for the first and final conference considered in the study (LREC 2010 and LREC 2012), and for the whole data corresponding to 10 conferences.

Table 1 gives some insight into the number of operations that were made for the first and last conference in the series, and for the analysis of the whole set. The number of corrections has been drastically reduced thanks to the new acquisition software regarding the names (45% to 8%) and the languages (9% to 1%) while it has remained stable for the Types and the Modalities. The number of corrections went down from 17% of the cells for LREC 2010 to 6% for LREC 2012. On the final complete table, more than 2,000 cells had to be manually modified, that is 13% of the cells considered in the *LRE Map*, which represents many working hours.

3.2. The ISCA Archive

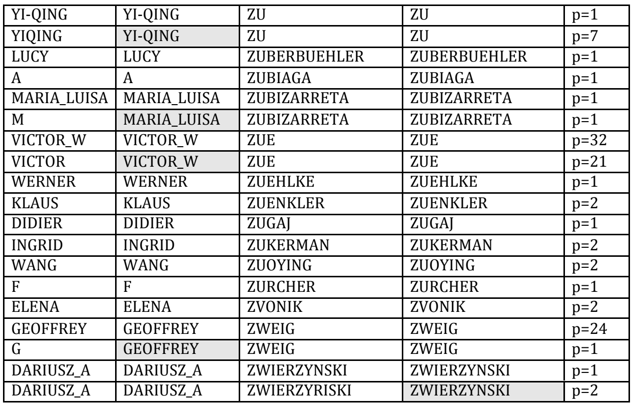
When the volume of data becomes very large, the hope of a comprehensive, ex post facto normalization becomes illusory. A study of the International Speech Communication Association (ISCA) Archive covering 25 annual conferences (ECST, Eurospeech, ICSLP and Interspeech) over a long duration (1987-2012) (J. Mariani et al., 2013) faced difficulties identifying the target information and requiring tedious cleaning on several occasions.

In addition to the detailed and substantiated look at the history of the ISCA community which provides valuable insights, this study enabled us to identify two important points to consider when doing community oriented bibliographic studies:

1. A cost-benefit perspective suggests that we focus on the data that have the greatest influence on survey goals. Normalizing the names of authors who published only one or two papers over 25 years has only a small effect despite the required effort. This is especially important given that more than half of the authors (about 8,000) published only one paper. In contrast, resolving the different names of an active author is important, because otherwise this person will not appear at the correct ranking, defeating one of the objectives for identifying the most active authors.
2. For a similar reason, using expert domain knowledge to ensure that the names of key people in the community are properly taken into account is unavoidable. Future work on multilingual enriched probabilistic parsing of person names is needed to increase the share of automated processing and improve referent identification performance.

3.2.1. Authors' Name Normalization

The data on the authors' names was obtained from the ISCA Archive metadata. After an automatic cleaning, the authors' names were listed in alphabetical order and a domain expert checked the top ranking authors according to frequency or domain importance and regarding the spelling of the family name and given name. The longest variant was retained as the reference. The substitution between the family name and the given name was also checked. From 16,892 person names, this processing reduced the list to 14,630 names, corresponding to an average ambiguity factor of 1.12 and a standard deviation of 0.448.

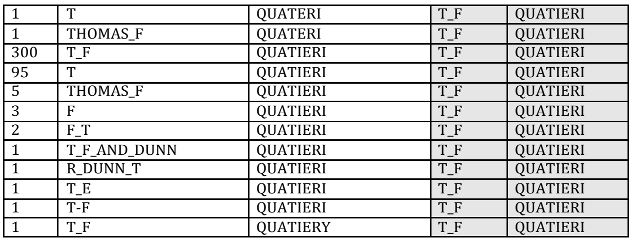


YI-QING	YI-QING	ZU	ZU	p=1
YIQING	YI-QING	ZU	ZU	p=7
LUCY	LUCY	ZUBERBUEHLER	ZUBERBUEHLER	p=1
A	A	ZUBIAGA	ZUBIAGA	p=1
MARIA_LUISA	MARIA_LUISA	ZUBIZARRETA	ZUBIZARRETA	p=1
M	MARIA_LUISA	ZUBIZARRETA	ZUBIZARRETA	p=1
VICTOR_W	VICTOR_W	ZUE	ZUE	p=32
VICTOR	VICTOR_W	ZUE	ZUE	p=21
WERNER	WERNER	ZUEHLKE	ZUEHLKE	p=1
KLAUS	KLAUS	ZUENKLER	ZUENKLER	p=2
DIDIER	DIDIER	ZUGAJ	ZUGAJ	p=1
INGRID	INGRID	ZUKERMAN	ZUKERMAN	p=2
WANG	WANG	ZUOYING	ZUOYING	p=2
F	F	ZURCHER	ZURCHER	p=1
ELENA	ELENA	ZVONIK	ZVONIK	p=2
GEOFFREY	GEOFFREY	ZWEIG	ZWEIG	p=24
G	GEOFFREY	ZWEIG	ZWEIG	p=1
DARIUSZ_A	DARIUSZ_A	ZWIERZYNSKI	ZWIERZYNSKI	p=1
DARIUSZ_A	DARIUSZ_A	ZWIERZYRISKI	ZWIERZYNSKI	p=2

Fig. 5. Example of cleaning authors' surnames (first and second columns) and names (third and fourth columns): the focus is on the most prolific authors (number of papers in fifth column), as merging variant wording may drastically change their ranking (see the case of Victor Zue / Victor W Zue, with 53 papers in total)

3.2.2. Cited Authors' Name Normalization

The same process was applied to the analysis of the authors cited in papers. The process is even more difficult, as the data is obtained through an automatic analysis of the papers content using the ParsCit software (I. Councill et al., 2008) and therefore contains segmentation errors. Also the number of cited authors is much larger than the number of ISCA Archive papers' authors. We first automatically cleaned the data by using the results of the former process on the authors' names, before conducting a manual cleaning focused on the most cited among the 51,144 cited authors.

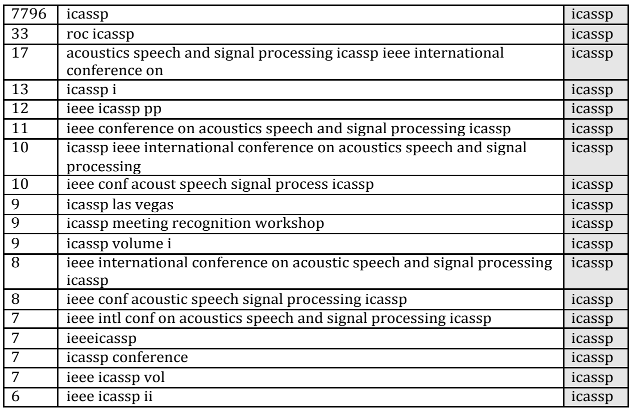


1	T	QUATERI	T_F	QUATIERI
1	THOMAS_F	QUATERI	T_F	QUATIERI
300	T_F	QUATIERI	T_F	QUATIERI
95	T	QUATIERI	T_F	QUATIERI
5	THOMAS_F	QUATIERI	T_F	QUATIERI
3	F	QUATIERI	T_F	QUATIERI
2	F_T	QUATIERI	T_F	QUATIERI
1	T_F_AND_DUNN	QUATIERI	T_F	QUATIERI
1	R_DUNN_T	QUATIERI	T_F	QUATIERI
1	T_E	QUATIERI	T_F	QUATIERI
1	T-F	QUATIERI	T_F	QUATIERI
1	T_F	QUATIERY	T_F	QUATIERI

Fig. 6. Example of cleaning the cited authors' surname (second and fourth columns) and name (third and fifth columns): here also the focus is put on the most cited authors (number of citations in first column), as merging variant wording may drastically change their ranking (from 300 to 412 citations for T.F. Quatieri)

3.2.3. Source of Reference Normalization

The sources of the citations were also extracted from the content by using ParsCit. They may refer to several categories of sources: conferences and workshops, journals or books. The cleaning was first conducted on a single year (2007). The resulting filter was then used for all the years, and the full data received a final review. Here also, only the most cited sources were considered (more than 5 citations), given the size of the data to consider (78,375 source citations).

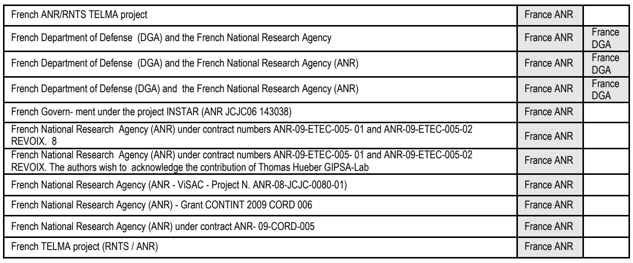


7796	icassp	icassp
33	roc icassp	icassp
17	acoustics speech and signal processing icassp ieee international conference on	icassp
13	icassp i	icassp
12	ieee icassp pp	icassp
11	ieee conference on acoustics speech and signal processing icassp	icassp
10	icassp ieee international conference on acoustics speech and signal processing	icassp
10	ieee conf acoust speech signal process icassp	icassp
9	icassp las vegas	icassp
9	icassp meeting recognition workshop	icassp
9	icassp volume i	icassp
8	ieee international conference on acoustic speech and signal processing icassp	icassp
8	ieee conf acoustic speech signal processing icassp	icassp
7	ieee intl conf on acoustics speech and signal processing icassp	icassp
7	ieeeicassp	icassp
7	icassp conference	icassp
7	ieee icassp vol	icassp
6	ieee icassp ii	icassp

Fig. 7. Example of cleaning the source of a reference (second and third columns): the focus is put on the most cited variants of a source (first column), as merging variant wordings change their ranking (from an initial 5662 to 9006 citations for ICASSP)

3.2.4. Funding Agency Normalization

The analysis of the Acknowledgements of the Funding sources on the papers contents also necessitated a manual cleaning. The nationality of each funding agency was introduced, and the spelling variants were harmonized in order to estimate the agencies and countries that are the most active in funding research on Spoken Language Processing.



French ANR/RNTS TELMA project	France ANR	
French Department of Defense (DGA) and the French National Research Agency	France ANR	France DGA
French Department of Defense (DGA) and the French National Research Agency (ANR)	France ANR	France DGA
French Department of Defense (DGA) and the French National Research Agency (ANR)	France ANR	France DGA
French Govern- ment under the project INSTAR (ANR JCJC06 143038)	France ANR	
French National Research Agency (ANR) under contract numbers ANR-09-ETEC-005- 01 and ANR-09-ETEC-005-02 REVOIX. 8	France ANR	
French National Research Agency (ANR) under contract numbers ANR-09-ETEC-005- 01 and ANR-09-ETEC-005-02 REVOIX. The authors wish to acknowledge the contribution of Thomas Hueber GIPSA-Lab	France ANR	
French National Research Agency (ANR - ViSAC - Project N. ANR-08-JCJC-0080-01)	France ANR	
French National Research Agency (ANR) - Grant CONTINT 2009 CORD 006	France ANR	
French National Research Agency (ANR) under contract ANR- 09-CORD-005	France ANR	
French TELMA project (RNTS / ANR)	France ANR	

Fig. 8. Example of cleaning the reference of Funding Agencies (first and second columns), with the case of mentioning several Funding Agencies (third column). The nationality of the Funding Agency is also mentioned.

3.3. The LDC Bibliography

The LDC Bibliography identifies and catalogs citations to papers mentioning Language Resources LDC publishes, including those that introduce or criticize the resource and those that describe efforts to build upon it, adding annotation or using it for technology development. Beneficiaries include the Language Resource producer and publisher who learn how and when the resource is used, what feedback users offer and, by extension, the impact of the resource. Potential Language Resource creators and users may also benefit by observing the dialog among other creators and users. LDC has identified more than 11,000 papers related to approximately 85% of more than 500 titles.

The principal goal of this effort is to identify technical papers that make some use of one or more LDC published Language Resources and to highlight the relationship between the papers and the resources. Given only the mentions of Language Resources as they appear in technical papers at publication time, the challenges we face focus on what we above labeled identification in synchrony and reference precision and what elsewhere are called entity normalization, mapping and co-reference among mentions of LDC resources. We have already seen, in the examples of the LRE Map and ISCA Archive, statistics on the amount of ambiguity introduced by the lack of normalized references. Here we will focus on specific issues that have arisen in our survey and that similarly challenge researchers attempting to understand the technical papers of their peers. In particular, we show that the variation in mentions of Language Resources is non-trivial and cannot be easily normalized.

We have already seen, in Figure 2 above, the variation in the name of the ACE corpora. In an even commoner example, several LRs are mentioned informally as the Treebank, Penn Treebank or PTB. These refer to the LDC publications Treebank-2 (LDC95T7) and Treebank-3 (LDC99T42). Because both include Wall Street Journal text, scholars frequently refer to that subset as the Wall Street Journal or WSJ corpus. In the literature surveyed to date, mentions take forms as varied as the 5K Wall Street Journal (WSJ), WSJ version of Penn Treebank, The Wall Street Journal (WSJ) task, Penn Treebank. Unfortunately mentions of WSJ may also refer to the text data in the TIPSTER (e.g. LDC93T3A) or North American News Text corpora (e.g. LDC2008T15), and the read versions in the so-called CSR corpora (e.g. LDC93S6A), as well as the syntactically annotated data in Treebank. Complicating matters, the CSR corpora, which contain readings of the Wall Street Journal, were released in two sets commonly known as WSJ0 and WSJ1. However, even these mentions are inadequate since each set was published as three corpora: two differing importantly in the microphone used and a third, complete collection. Thus mentions in the literature of WSJ0, WSJ1 or worse, the Wall Street Journal Corpus fail to provide an adequate pointer to the data used for a researcher hoping to replicate or compare prior studies. The reference problems are not limited to older corpora either. More recent mentions of the Web 1T 5-gram corpus Version 1 (LDC2006T13) appear in the literature as Google Web 1T n-gram, Google 5-gram and Google Ngrams.

Readers may have difficulty determining from such mentions which corpus was in fact used. LDC domain experts resolve these ambiguities through their knowledge of the LRs themselves and by cross-checking against the LRs known to be licensed to the paper's author. However, knowing that no reader will have access to such information and also knowing that information extraction technologies are imperfect and expensive to build and maintain, it strikes us as wasteful to continue to try to address such entity and relation normalization problems ex post facto.

4. Conclusion

Each of the three studies that are mentioned in this paper has resulted in interesting results allowing for a better understanding of our field of research. They have benefited from a large amount of data and an extensive set of tools. However, they required a huge amount of manual work in order to normalize and identify the information related to the various parameters that were taken into account in the analysis, with the help of referential data whenever they exist.

This tedious activity has clearly demonstrated the need for normalizing, assigning and using persistent and unique identifiers for various entities: authors names, affiliations, gender and nationality, conferences, journals, books, papers references, Funding Agencies, Language Resources names, types and modalities, languages, etc. which would necessitate an international coordination, in order to be able to draw reliable statistics from existing scientific literature, a key asset for measuring science progress.

5. Acknowledgements

This paper results from studies conducted within the EC projects FLaReNet (Grant Agreement ECP-2007-LANG-617001) and T4ME (Grant Agreement no.: 249119)).

The authors wish to thank all the people who participated in those studies, especially Nicoletta Calzolari, Riccardo del Gratta, Irene Russo and Claudia Soria (CNR-ILC "A. Zampolli"), and Khalid Choukri (ELRA-ELDA) for fruitful discussions about the *LRE Map* content and Denise DiPersio, Eleftheria Ahtaridis and Abby Graham for discussions of the LDC Bibliography.

6. References

ACL (2012) Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, ACL 2012, Jeju, July 10, 2012, ISBN 978-1-937284-29-9

Ahtaridis, Eleftheria; Cieri, Christopher and DiPersio, Denise (2012) LDC Language Resource Papers: Building a Bibliographic Database, In Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, 23-25 May 2012.

Calzolari, Nicoletta; Del Gratta, Riccardo; Francopoulo, Gil; Mariani, Joseph; Rubino, Francesco; Russo, Irene and Soria, Claudia (2012) The LRE Map. Harmonising Community Descriptions of Resources, In Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, 23-25 May 2012

Calzolari, Nicoletta and Mariani, Joseph (2012) A journey from the LRE Map to the Language Matrices and the Language Resource Impact Factor (LRIF), Coco-FLaRe Workshop "Reinforcing International Collaboration on LRE", LREC 2012, 26 May 2012

Cambon-Thomsen, A., Thorisson, G. A. and Mabile, L. (2011) The role of a bioresource research impact factor as an incentive to share human bioresources, Nature Genetics, 43, 503–504, 26 may 2011

Choukri, Khalid; Arranz, Victoria; Hamon, Olivier and Park, Jungyeul (2012) Using the International Standard Language Resource Number: Practical and Technical Aspects, In Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, 23-25 May 2012

Councill, Isaac G.; Giles, C. Lee; Kan, Min-Yen (2008) ParsCit: An open-source CRF reference string parsing package. In Proceedings of the Language Resources and Evaluation Conference (LREC 2008), Marrakesh, Morocco, 28-30 May 2008

Labropoulou, Penny; Cieri, Christopher and Gavriilidou, Maria (2104) Developing a Framework for Describing Relations among Language Resources, In Proceedings of the Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, 26-31 May 2014.

Lewis, Paul, ed. (2009) Ethnologue: Languages of the World, Sixteenth Edition, 2009, 1,248 pp., SIL International, ISBN 978-1-55671-216-6

Mariani, Joseph; Paroubek, Patrick; Francopoulo, Gil and Delaborde, Marine (2013) Rediscovering 25 years of Discoveries in Spoken Language Processing: a Preliminary Analysis of the ISCA Archive, Interspeech'2013, Lyon, 26-29 August 2013

Soria, Claudia and Mariani, Joseph (2013) Searching LTs for minority languages, Workshop TALaRE Traitement Automatique des Langues Régionales de France et d'Europe, Sables d'Olonne, 21 juin 2013