



LGPLLR : an open source license for NLP (Natural Language Processing)

Sébastien Paumier

Université Paris-Est Marne-la-Vallée



Linguistic data

- data meant to be used by programs: different from electronic versions of things like human dictionaries
- a continuum between raw data like word lists and highly-structured data like tree banks and word nets:
 - electronic dictionaries
 - syntactic grammars
 - annotated corpora
 - etc.



Linguistic data

- anyone has some native expertise about language: only programmers can fix bugs, but anyone can see that *ptesident* is a typing error
- a bad conclusion: anyone can deal with linguistic data, it is not a serious work
- an a good one: many people could contribute to make data better



No serious work ?

- anyone that has built linguistic resources knows how difficult it can be, even if the result looks obvious
- example: country names
 - how to script them in your alphabet ?
 - what is the official definition of *country* ?
 - how to deal with names with a determiner like *The United States of America* ?
 - how about variants like U.S.A ?
 - etc.



No serious work ?

- the amount of work should not be underestimated
- a good question to ask: is the work important enough to be protected ?

	computer science	linguistic data
no	integer linked lists	list of English 1-letter words
yes	automata library	syntactic lexicon

- if the answer is yes, you need a license for your data



Cooperative work

- linguistic data may be easier to fix and extend than programs, but available manpower is useless if data are not modifiable
- example: EuroWordNet
 - people cannot commit their modifications
 - same waste of time as for non-free software
- solution: allow people to modify the data, with an appropriate license



LGPLLR

- derived from the LGPL, adapted for linguistic data
- *linguistic resource*: collection of data about language prepared so as to be used with application programs



LGPLLR

- *a work based on the linguistic resource*: either the linguistic resource or any derivative work containing it or a portion of it, either verbatim or with modifications (including translation)
- *legible form*: the preferred form of the resource for making modifications to it



Your rights

- anyone can redistribute the data modified or as is, with a copy of the license
- the modified work must be a linguistic resource:
 - subset of adjectives is a linguistic resource derived from a lexicon
 - a letter frequency table is not: such a resource is not covered by the license



Your obligations

- modified files must carry prominent notices stating that you changed them and the date of any change
- moreover, it is strongly suggested that you explain your modifications (one word more or less can have a strong meaning)
- you must redistribute the machine-readable legible form of the linguistic resources
 - a printed appendix in a book is not sufficient



If you write programs...

- a program that contains no derivative of any portion of the data, but is designed to work with is called *a work that uses the linguistic resource*, and falls outside the scope of the LGPL
- however, combining such a program with data is considered as a derivative of the linguistic resource



If you write programs...

- you must use a suitable mechanism for combining with the linguistic resource
- a suitable mechanism is one that will operate properly with a modified version of the linguistic resource
- if your package includes an encrypted version of the linguistic resources, it must contain any data and utility programs needed to rebuild this encrypted version from the original data



Some works using LGPLLLR

- Tables du lexique-grammaire du français
- Lefff: lexique des formes fléchies du français
- Prolex: lexique de noms propres
- DicoValence: dictionnaire de valence des verbes français
- HPSG FRoG French Resource Grammar: An HPSG grammar of French developed with the LKB platform (parsing and generation)



Some works using LGPLLLR

- SynLex: syntactic lexicons derived from lexicon-grammar tables
- Sanskrit linguistic resources
- Spanish Resource Grammar
- GerTT: grammar fragment of German in TT-MCTAG
- LexSchem: lexique de schémas de sous-catégorisation des les verbes français automatiquement acquis à partir de corpus



Some works using LGPLLLR

- GG: an HPSG for German
- Les Verbes français de J. Dubois et F. Dubois-Charlier
- Locutions en Français de J. Dubois et F. Dubois-Charlier
- all dictionaries and grammars distributed with Unitex



Conclusion

- linguistic resources with restricted distribution policies suffer from the same problems that non-free software
- freely accessible data are good
- freely modifiable data are better
- share your work and make other's better:

use LGPL

- <http://igm.univ-mlv.fr/~unitex/>