Folheador: browsing through Portuguese semantic relations

Hugo Gonçalo Oliveira
CISUC, University of Coimbra
Portugal
hroliv@dei.uc.pt

Hernani Costa
FCCN, Linguateca &
CISUC, University of Coimbra
Portugal
hpcosta@dei.uc.pt

Diana Santos
FCCN, Linguateca &
University of Oslo
Norway
d.s.m.santos@ilos.uio.no

Abstract

This paper presents Folheador, an online service for browsing through Portuguese semantic relations, acquired from different sources. Besides facilitating the exploration of Portuguese lexical knowledge bases, Folheador is connected to services that access Portuguese corpora, which provide authentic examples of the semantic relations in context.

1 Introduction

Lexical knowledge bases (LKBs) hold information about the words of a language and their interactions, according to their possible meanings. They are typically structured on word senses, which may be connected by means of semantic relations. Besides important resources for language studies, LKBs are key resources in the achievement of natural language processing tasks, such as word sense disambiguation (see e.g. Agirre et al. (2009)) or question answering (see e.g. Pasca and Harabagiu (2001)).

Regarding the complexity of most knowledge bases, their data formats are generally not suited for being read by humans. User interfaces have thus been developed for providing easier ways of exploring the knowledge base and assessing its contents. For instance, for LKBs, in addition to information on words and semantic relations, it is important that these interfaces provide usage examples where semantic relations hold, or at least where related words co-occur.

In this paper, we present Folheador[1], an online browser for Portuguese LKBs. Besides an interface for navigating through semantic relations acquired from different sources, Folheador is linked to two services that provide access to Portuguese corpora, thus allowing observation of related words co-occurring in authentic contexts of use, some of them even evaluated by humans.

After introducing several well-known LKBs and their interfaces, we present Folheador and its main features, also detailing the contents of the knowledge base currently browseable through this interface, which contains information acquired from public domain lexical resources of Portuguese. Then, before concluding, we discuss additional features planned for the future.

2 Related Work

Here, we mention a few interfaces that ease the exploration of well-known knowledge bases. Regarding the knowledge base structure, some of the interfaces are significantly different.

Princeton WordNet (Fellbaum, 1998) is the most widely used LKB to date. In addition to other alternatives, the creators of WordNet provide online access to their resource through the WordNet Search interface (Princeton University, 2010)[2]. As WordNet is structured around synsets (groups of synonymous lexical items), querying for a word prompts all synsets containing that word to be presented. For each synset, its part-of-speech (PoS), a gloss and a usage example are provided. Synsets can also be expanded to access the semantic relations they are involved in.

As a resource also organised in synsets, the

[1] See http://www.linguateca.pt/Folheador/

[2] http://wordnetweb.princeton.edu/perl/webwn

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 35–40,
Avignon, France, April 23 - 27 2012.

Brazilian Portuguese thesaurus TeP[3] has a similar interface (Maziero et al., 2008). Nevertheless, since TeP does not contain relations besides antonymy, its interface is simpler and provides only the synsets containing a queried word and their part-of-speech.

MindNet (Vanderwende et al., 2005) is a LKB extracted automatically, mainly from dictionaries, and structured on semantic relations connecting word senses to words. Its authors provide MNEX[4], an online interface for MindNet. After querying for a pair of words, MNEX provides all the semantic relation paths between them, established by a set of links that connect directly or indirectly one word to another. It is also possible to view the definitions that originated the path.

FrameNet (Baker et al., 1998) is a manually built knowledge base structured on semantic frames that describe objects, states or events. There are several means for exploring FrameNet easily, including FrameSQL (Sato, 2003)[5], which allows searching for frames, lexical units and relations in an integrated interface, and FrameGrapher[6], a graphical interface for the visualization of frame relations. For each frame, in both interfaces, a textual definition, annotated sentences of the frame elements, lists of the frame relations, and lists with the lexical units in the frame are provided.

ReVerb (Fader et al., 2011) is a Web-scale information extraction system that automatically acquires binary relations from text. Using ReVerb Search[7], a web interface for ReVerb extractions, it is possible to obtain sets of relational triples where the predicate and/or the arguments contain given strings. Regarding that each of the former is optional, it is possible, for instance, to search for all triples with the predicate *loves* and first argument *Portuguese*. Search results include the matching triples, organised according to the name of the predicate, as well as the number of times each triple was extracted. The sentences where each triple was extracted from are as well provided.

Finally, Visual Thesaurus (Huiping et al., 2006)[8] is a proprietary graphical interface that provides an alternative way of exploring a knowledge base structured on word senses, synonymy, antonymy and hypernymy relations. It presents a graph centered on a queried word, connected to its senses, as well as semantic relations between the senses and other words. Nodes and edges have a different color or look, respectively according to the PoS of the sense or to the type of semantic relation. If a word is clicked, a new graph, centered on that word, is drawn.

3 Folheador

Folheador, in figure 2, is an online service for browsing through instances of semantic relations, represented as relational triples.

Folheador was originally designed as an interface for PAPEL (Gonçalo Oliveira et al., 2010), a public domain lexical-semantic network, automatically extracted from a proprietary dictionary. It was soon expanded to other (public) resources for Portuguese as well (see Santos et al. (2010) for an overview of Portuguese LKBs).

The current version of Folheador browses through a LKB that, besides PAPEL, integrates semantic triples from the following sources: (i) synonymy acquired from two handcrafted thesauri of Portuguese[9], TeP (Dias-Da-Silva and de Moraes, 2003; da Silva et al., 2002) and OpenThesaurus.PT[10]; (ii) relations extracted automatically in the scope of the project Onto.PT (Gonçalo Oliveira and Gomes, 2010; Gonçalo Oliveira et al., 2011), which include triples extracted from Wiktionary.PT[11], and from Dicionário Aberto (Simões and Farinha, 2011), both public domain dictionaries.

Underlying relation triples in Folheador are thus in the form *x* RELATED-TO *y*, where *x* and *y* are lexical items and RELATED-TO is a predicate. Their interpretation is as follows: one sense of *x* is related to one sense of *y*, by means of a relation whose type is identified by RELATED-TO.

[3] http://www.nilc.icmc.usp.br/tep2

[4] http://stratus.research.microsoft.com/mnex/

[5] http://framenet2.icsi.berkeley.edu/frameSQL/fn2_15/notes/

[6] https://framenet.icsi.berkeley.edu/fndrupal/FrameGrapher

[7] http://www.cs.washington.edu/research/textrunner/reverbdemo.html

[8] http://www.visualthesaurus.com/

[9] We converted the thesauri to triples *x* synonym-of *y*, where *x* and *y* are lexical items in the same synset.

[10] http://openthesaurus.caixamagica.pt/

[11] http://pt.wiktionary.org/

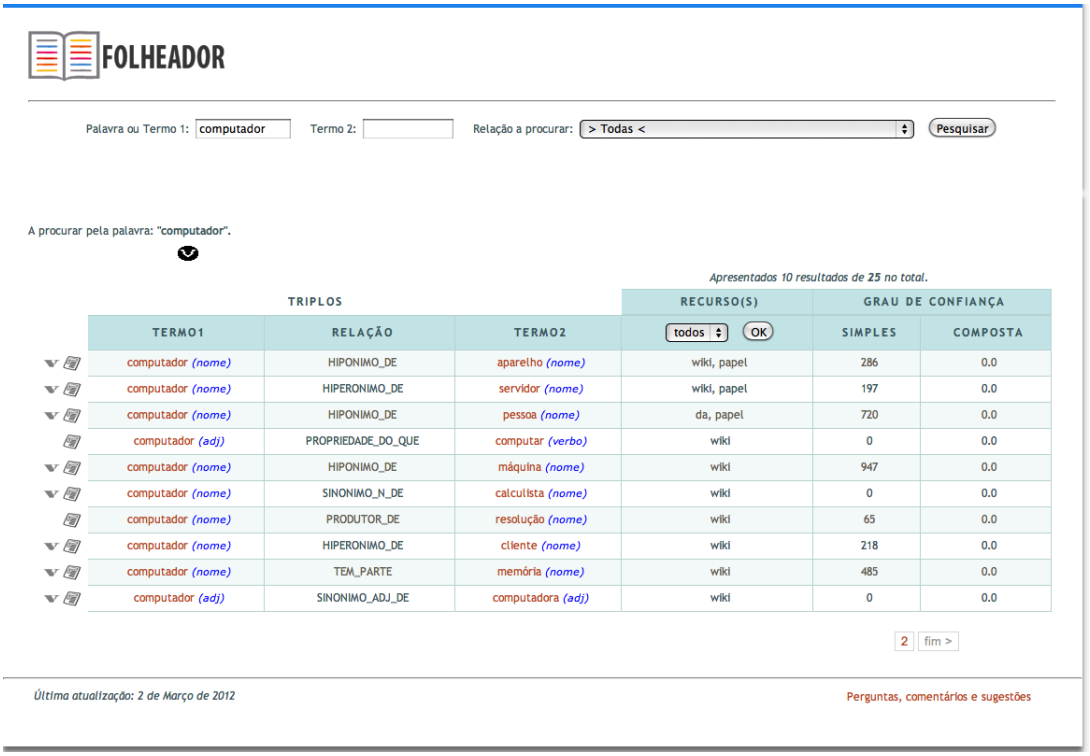

Figure 1: Folheador's interface.

3.1 Navigation

It is possible to use Folheador for searching for all relations with one, two, or no fixed arguments, and one or no types (relation names). Combining these options, Folheador can be used, for instance, to obtain: all lexical items related to a particular item; all relations between two lexical items; or a sample of relations involving a particular type.

The matching triples are listed and may be filtered according to the resource they were extracted from. For each triple, the PoS of the arguments is shown, as well as a list with the identification of the resources from where it was acquired. The arguments of each triple are also links that make navigation easier. When clicked, Folheador behaves the same way as if it had been queried with the clicked word as argument. Also, since the queried lexical item may occur in the first or in the second argument of a triple, when it occurs in the second, Folheador inverts the relation, so that the item appears always as the first argument. Therefore, there is no need to store both the direct and the inverse triples.

Consider the example in figure 2: it shows the triples retrieved after searching for the word *computador* (computer, in English). In most of the retrieved triples, *computador* is a noun (e.g. *computador* HIPONIMO_DE *máquina*), but there are relations where it is an adjective (e.g. *computador* PROPRIEDADE_DO_QUE *computar*). Moreover, as hypernymy relations are stored in the form *x* HIPERONIMO_DE *y*, some of the triples presented, such as *computador* HIPONIMO_DE *máquina* and *computador* HIPONIMO_DE *aparelho*, have been inverted on the fly.

Furthermore, for each triple, Folheador presents: a confidence value based on the mere co-occurrence of the words in corpora; and another based on the co-occurrence of the related words instantiating discriminating patterns of the particular relation.

3.2 Graph visualization

Currently, Folheador contains a very simple visualization tool, which draws the semantic relation graph established by the search results in a page, as in figure 3.2. In the future, we aim to provide an alternative for navigation based on textual links, which would be made through the graph.

3.3 The use of corpora

One of the problems of most lexical resources is that they do not integrate or contain frequency in-

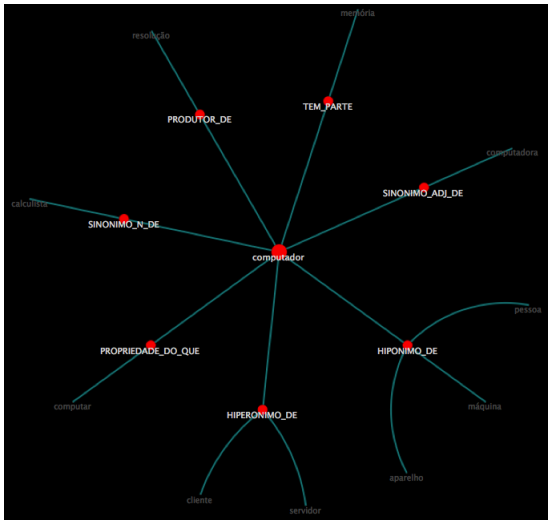

Figure 2: Graph for the results in figure 2.

formation. This is especially true when one is not simply listing words but going deeper into meaning, and listing semantic properties like word senses or relationships between senses.

So, a list of relations among words can conflate a number of highly specialized and obsolete words (or word senses) that co-occur with important and productive relations in everyday use, which is not a good thing for human and automatic users alike. On the other hand, using corpora allows one to add frequency information to both participants in the relation and the triples themselves, and thus provide another axis to the description of words.

In addition, it is always interesting to observe language use in context, especially in cases where the user is not sure whether the relation is correct or still in use (and the user can and should be fairly suspicious when s/he is browsing automatically compiled information). A corpus check therefore provides illustration, and confirmation, to a user facing an unusual or surprising relation, in addition to evaluation data for the relation curator or lexicographer. If these checks have been done before by a set of human beings (as is the case of VARRA (Freitas et al., forthcomming)), one can have much more confidence on the data browsed, something that is important for users.

Having this in mind, besides allowing to query for stored relational triples, Folheador is connected to AC/DC (Santos and Bick, 2000; Santos, 2011), an online service that provides access to a large set of Portuguese corpora. In just one click, it is possible to query for all the sentences in the AC/DC corpora connecting the arguments of a retrieved triple. Figure 3.3 shows some of the results for the words *computador* (computer) and *aparelho* (apparatus). While some of the returned sentences might contain the related words co-occurring almost by chance or without a clear semantic relation, other sentences validate the triple (e.g. sentence *par=saude16727* in figure 3.3). Sometimes, the sentences might as well invalidate the triple.

Furthermore, for some of the relation types, it is possible to connect to another online service, VARRA (Freitas et al., forthcomming), which is based on a set of patterns that express some of the relation types, in corpora text. After clicking on the VARRA link, this service is queried for occurrences of the corresponding triple in AC/DC. The presented sentences (a subset of those returned by the previous service) will thus contain the related words connected by a discriminating pattern for the relation they hold. Figure 3.3 shows two sentences returned for the relation *computador* HIPONIMO_DE *máquina*.

These patterns, as those proposed by Hearst (1992) and used in many projects since, may not be 100% reliable. So, VARRA was designed to allow human users to classify the sentences according to whether the latter validate the relation, are just compatible with it, or not even that.

In fact, people do not usually write definitions, especially when using common sense terms in ordinary discourse. Thus, co-occurrence of semantically-related terms frequently indicates a particular relation only implicitly. The choice of assessing sentences as good validators of a semantic relation is related to the task of automatically finding good illustrative examples for dictionaries, which is a surprisingly complex task (Rychlý et al., 2008).

This kind of information, amassed with the help of VARRA, is much more difficult to create, but is of great value to Folheador, since it provides good illustrative contexts for the related lexical items.

4 Further work and concluding remarks

We have shown that, as it is, Folheador is very useful, as it enables to browse for triples with fixed arguments, it identifies the source of the triples, and, in one click, it provides real sentences

par=2530: Ela trazia irregularmente do Paraguai computadores, **aparelhos** eletrônicos e uísque .
par=2548: Em outubro, Wanderlei usou cerca de r $ 15 mil que Márcia havia juntado com os seus contrabandos para comprar duas televisões, dois videocassetes, um **aparelho** de som, uma filmadora e um computador .
: · Para você que quer ter uma coleção de músicas para seu computador ou mesmo para ouvir no seu novo **aparelho** de MP3, não perca essa oportunidade .
: Usando o CyberTracker, um software com o qual os ecologistas podem registrar suas observações em campo usando computadores portáteis conectados a **aparelhos** de posicionamento global (GPS) , os rastreadores puderam reunir dados que comprovam a degradação da população local da espécie .
: · Para você que quer ter uma coleção de músicas para seu computador ou mesmo para ouvir no seu novo **aparelho** de MP3, não perca essa oportunidade .
: Segundo a pesquisa, 16,6 % dos domicílios brasileiros têm computadores de mesa, contra 95,7 % que têm **aparelhos** de TV .
par=49126: O **aparelho** está equipado com modernos instrumentos de telecomunicações, primeiros-socorros, páraquedas e computador .
par=saude16727: Os avanços da ecografia, enquanto tecnologia, resultam da evolução da Infor mática, afinal, estes **aparelhos** são computadores que analisam o som e a imagem .

Figure 3: AC/DC: some sentences returned for the related words *computador* and *aparelho*.

Relação	Procura	Exemplo
máquina HIPERONIMO_DE computador	padrões usados	*par=Mais--94a-2*: E também de ensinar **máquinas** como **computadores** a identificarem 'ses objetos . (NSC)
máquina HIPERONIMO_DE computador	padrões usados	*par=ext328388-soc-95a-2*: **Máquinas** como os **computadores**, os faxes e os videofones devem poder comunicar entre si sem falhas, o que supõe um trabalho de programação importante . (CP)

Figure 4: VARRA: sentences that exemplify the relation *computador* hyponym-of *máquina*.

where related lexical items co-occur. Still, we are planning to implement new basic features, such as the suggestion of words, when the searched word is not in the LKB. Also, while currently Folheador only directly connects to AC/DC and VARRA, in order to increase its usability, we plan to connect it automatically to online definitions and other services available on the Web. We intend as well to crosslink Folheador from the AC/DC interface, in the sense that one can invoke Folheador also by just one click (Santos, forthcomming).

Currently, Folheador gives access to 169,385 lexical items: 93,612 nouns, 38,409 verbs, 33,497 adjectives and 3,867 adverbs, in a total of 722,589 triples, and it can browse through the following types of semantic relations: synonymy, hypernymy, part-of, member-of, causation, producer-of, purpose-of, place-of, and property-of. However, as the underlying resources, especially the ones created automatically, will continue to be updated, one important challenge is to create a service that does not get outdated, by accompanying the progress of these resources, ideally doing an automatic update every month. Furthermore, we believe that quantitative studies on the comparison and the aggregation of the integrated resources should be made, deeper than what is presented in Gonçalo Oliveira et al. (2011).

We would like to end by emphasizing that we are aware that the proper interpretation of the semantic relations may vary in the different resources, even disregarding possible mistakes in the automatic harvesting. It is enough to consider the (regular morphological) relation between a verb and an adjective/noun ended in *-dor* in Portuguese (and which can be paraphrased by one who Vs). For instance, in relations such as {sofrer - sofredor}, {correr - corredor}, {roer - roedor}, the kind of verb defines the kind of temporal relation conveyed: a rodent is essentially *roendo*, while a *sofredor* (sufferer) suffers hopefully in a particular situation and can stop suffering, and a *corredor* (runner) runs as job or as role.

The source code of Folheador is open source[12], so it may be used by other authors to explore their knowledge bases. Technical information about Folheador may be found in Costa (2011).

Acknowledgements

Folheador was developed under the scope of Linguateca, throughout the years jointly funded by the Portuguese Government, the European Union (FEDER and FSE), UMIC, FCCN and FCT. Hugo Gonçalo Oliveira is supported by the FCT grant SFRH/BD/44955/2008 co-funded by FSE.

References

Eneko Agirre, Oier Lopez De Lacalle, and Aitor Soroa. 2009. Knowledge-based WSD on specific domains: performing better than generic supervised WSD. In *Proceedings of 21st Interna-*

[12] Available from `http://code.google.com/p/folheador/`

tional Joint Conference on Artifical Intelligence, IJCAI'09, pages 1501–1506, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th international conference on Computational linguistics*, pages 86–90, Morristown, NJ, USA. ACL Press.

Hernani Costa. 2011. O desenho do novo Folheador. Technical report, Linguateca.

Bento C. Dias da Silva, Mirna F. de Oliveira, and Helio R. de Moraes. 2002. Groundwork for the Development of the Brazilian Portuguese Wordnet. In Nuno Mamede and Elisabete Ranchhod, editors, *Advances in Natural Language Processing (PorTAL 2002)*, LNAI, pages 189–196, Berlin/Heidelberg. Springer.

Bento Carlos Dias-Da-Silva and Helio Roberto de Moraes. 2003. A construção de um thesaurus eletrônico para o português do Brasil. *ALFA*, 47(2):101–115.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference of Empirical Methods in Natural Language Processing*, EMNLP '11, Edinburgh, Scotland, UK.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Cláudia Freitas, Diana Santos, Hugo Gonçalo Oliveira, and Violeta Quental. forthcomming. VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC. In *Atas do IX Encontro de Linguística de Corpus*, ELC 2010.

Hugo Gonçalo Oliveira and Paulo Gomes. 2010. Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)*, pages 199–211. IOS Press.

Hugo Gonçalo Oliveira, Diana Santos, and Paulo Gomes. 2010. Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. *Linguamática*, 2(1):77–93.

Hugo Gonçalo Oliveira, Leticia Antón Pérez, Hernani Costa, and Paulo Gomes. 2011. Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. *Linguamática*, 3(2):23–38.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of 14th Conference on Computational Linguistics*, pages 539–545, Morristown, NJ, USA. ACL Press.

Du Huiping, He Lin, and Hou Hanqing. 2006. Thinkmap visual thesaurus:a new kind of knowledge organization system. *Library Journal*, 12.

Erick G. Maziero, Thiago A. S. Pardo, Ariani Di Felippo, and Bento C. Dias-da-Silva. 2008. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In *VI Workshop em Tecnologia da Informação e da Linguagem Humana*, TIL 2008, pages 390–392.

Marius Pasca and Sanda M. Harabagiu. 2001. The informative role of WordNet in open-domain question answering. In *Proceedings of NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pages 138–143, Pittsburgh, USA.

Princeton University. 2010. Princeton university "About Wordnet". http://wordnet.princeton.edu.

Pavel Rychlý, Miloš Husák, Adam Kilgarriff, Michael Rundell, and Katy McAdam. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress*, pages 425–432, Barcelona. Institut Universitari de Lingüística Aplicada.

Diana Santos and Eckhard Bick. 2000. Providing Internet access to Portuguese corpora: the AC/DC project. In *Proceedings of 2nd International Conference on Language Resources and Evaluation*, LREC'2000, pages 205–210. ELRA.

Diana Santos, Anabela Barreiro, Cláudia Freitas, Hugo Gonçalo Oliveira, José Carlos Medeiros, Luís Costa, Paulo Gomes, and Rosário Silva. 2010. Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL. In A. M. Brito, F. Silva, J. Veloso, and A. Fiéis, editors, *Textos seleccionados. XXV Encontro Nacional da Associação Portuguesa de Linguística*, pages 681–700. APL.

Diana Santos. 2011. Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. *OSLa: Oslo Studies in Language*, 3(2):113–128. Volume edited by J.B.Johannessen, Language variation infrastructure.

Diana Santos. forthcomming. Corpora at linguateca: vision and roads taken. In Tony Berber Sardinha and Telma S ao Bento Ferreira, editors, *Working with Portuguese corpora*.

Hiroaki Sato. 2003. FrameSQL: A software tool for FrameNet. In *Proceedings of Asialex 2003*, pages 251–258, Tokyo. Asian Association of Lexicography, Asian Association of Lexicography.

Alberto Simões and Rita Farinha. 2011. Dicionário Aberto: Um novo recurso para PLN. *Vice-Versa*, pages 159–171.

Lucy Vanderwende, Gary Kacmarcik, Hisami Suzuki, and Arul Menezes. 2005. Mindnet: An automatically-created lexical resource. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pages 8–9, Vancouver, British Columbia, Canada. ACL Press.