Supporting Knowledge Discovery for Biodiversity

Manuel Vilares*[,a], Milagros Fernández[a], Adrían Blanco[a]

[a]*Department of Computer Science, University of Vigo*
Campus As Lagoas s/n, 32004 – Ourense, Spain

Abstract

A proposal for text mining as a support for knowledge discovery on biological descriptions is introduced. Our aim is both to sustain the curation of databases and to offer an alternative representation frame for accessing information in the biodiversity domain. We works on raw texts with minimum human intervention, applying natural language processing to integrate linguistic and domain knowledge in a mathematical model that makes it possible to capture concepts and relationships between them in a computable form, using conceptual graphs. This provides a reasoning basis for determining semantic disjointedness or subsumption, as well as sub and super-concept relationships.

Key words: knowledge discovery, natural language processing, text mining

Notice:. **This is the authors version of a work that was accepted for publication in Data & Knowledge Engineering. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Data & Knowledge Engineering, 100(A):34-53, 2015. DOI 10.1016/j.datak.2015.08.002**

1. Introduction

Biodiversity description provides basic understanding for decision-making about conservation and sustainable use, affecting a wide range of sectors of both human and economic importance, such as the chemical and agri-food industries. This supports the interest of *taxonomy*, the science of describing, naming and classifying living organisms in an ordered system of *taxa*, largely considered to be unfashionable. So, it is often supposed that DNA barcoding is the ultimate solution to taxa identification, when in fact the arguments in its favour are illusory even for its proponents (Goldstein and DeSalle, 2011). People also assume that identifying species is a straightforward and

*Corresponding author: tel. +34 988 387280, fax +34 988 387001.
Email addresses: vilares@uvigo.es (Manuel Vilares), mfgavilanes@uvigo.es (Milagros Fernández), adbgonzalez@uvigo.es (Adrían Blanco)

low cost task, but that is far from being the case. We have only to realise that about 2 million species have been documented so far, which means that 80-90% of life is still to be discovered (Wilson, 2003). Furthermore, taxonomists not only delimitate taxa annotating descriptions for new species but also continually refine and review existing ones. Given that the data are distributed across thousands of journals and are provided by different researchers possibly using different vocabularies and methodologies, pursuing particular goals and working under varying spatio-temporal frames (Daltio and Medeiros, 2008), taxonomy becomes a complex task of knowledge management. As a result, there is a pressing need for capturing all this information in a way that is semantic, extensible and broadly accessible, which naturally leads us to ontologies (Smith et al., 2007). Unfortunately, their generation is too labor-intensive and time-consuming to ever be be fully automated, the process relying on qualified experts also known as *curators*.

With respect to access to information, not much has changed since the days of Linnaeus (Ereshefsky, 2007), who proposed the use of decision trees to identify taxa. Baptized as *keys*, their generation is a task reserved to curators and, while can be integrated in ontologies using `"is-a"` links, they have their weaknesses (Taylor, 95). So, some characteristics may have been omitted in the key due to error or absence at the moment the description was made, such as fruit properties, making them difficult to use. Also, identification can follow an unsuccessful path through the key, either due to the atypical nature of the specimen or to an error in determining whether it meets a decision criterion. This requires a return to the correct path, which is not a trivial task, especially for non-expert users.

We can then conclude that textual descriptions are not only of interest for database curation, but also for identifying species regardless of the user's expertise. Thus, *knowledge discovery* (KD) facilities are increasingly neccessary to support manual work, which justifies the interest in *text mining* (TM) techniques to perform *knowledge extraction* (KE) tasks (Deans et al., 2012).

2. The state-of-the-art

Roughly speaking, TM refers to the process of deriving new knowledge from text, which is often interpreted as comprising three major tasks, namely *information retrieval* (IR), *information extraction* and *data mining*. We can distinguish two approaches: *co-occurrence* and *natural language processing* (NLP) based TM.

2.1. Co-occurrence-based text-mining

Associations between terms are inferred on the assumption that when present in the same sentence or abstract they are related, following a semantic model known as *bag-of-words* (BOW) (Harris, 1954). The meaning of a text is represented by the multiset of its terms assuming full independence between them. Algebraic (Salton et al., 1975) and probabilistic (Maron and Kuhns, 1960) approaches are mainly used but, since little attention is paid to the linguistic structure, the type of association is neither identified nor negation dealt with, and thus non-meaningful relationships can arise. To minimize the latter, authors apply weighting criteria to rank the associations, such as

term frequency (TF), *inverse document frequency* (IDF) and document length (Salton and Buckley, 1988). All the above limits the potential interest of this approach for exploratory TM tasks, it now being used more as a baseline method against which others are compared (Zweigenbaum et al., 2007).

2.2. NLP*-based text-mining*

Co-occurrence provides recall, but we need access to a wealth of background knowledge in order to improve precision (Jensen et al., 2006). This places us within the context of NLP techniques, where syntactic and semantic analyses are combined with morphological and lexical variation through *part-of-speech tagging* (POST), to reveal relationships.

2.2.1. Syntactic modelling

We distinguish three models on the basis of the strategy applied to represent the meaning: *semantic*, *constraint-based logical* and *dependency grammars* (DG). The former (Brown and Burton, 1975) fills semantic templates according to sentence patterns. Most proposals (Shah et al., 2005) rely on *context-free grammars*, including *regular* ones (Miotto et al., 2005). The lack of contextual sensitivity favours non-determinism, often reduced by the consideration of restrictive sublanguages (Friedman et al., 2001) or domain-specific heuristics (Sekimizu et al., 1998), which do not go to the heart of the problem and impose the use of specialised grammars. This justifies the interest in formalisms with richly structured lexicons, such as *head-driven phrase structure grammars* (Creary and Pollard, 1985), although their applicability is questionable when the elements involved in the relevant constructions are not definable in strongly configurational terms (Levine, 2006). Alternatively, *mildly context-sensitive grammars* (MCSG) have acquired popularity in the sphere of NLP (Nesson et al., 2010) due not only to their lexical sensitivity (Schabes et al., 1988), but also to their capacity to deal with certain cross- and long-distance dependencies in polynomial time and space through the treatment of non-determinism in dynamic programming (de la Clergerie, 2010). This makes it possible to save all parses, postponing the resolution of ambiguities to a semantic stage.

Logical approaches look for the expressiveness of *first-order logic* (FOL) through rules associating predicates and semantic constraints by unification, providing parsing as deduction. The most popular one (Mungall, 2004; Taylor, 95) refers to *definite clause grammars* (Pereira and Warren, 1980), which pose problems of maintenance due to the fixed arity in predicates, meaning thatif we wish to extend a grammar each rule must be changed.

Both semantic and constraint-based logical grammars serve as a kernel for *phrase structure parsers*, which break sentences into constituents and can lead to complex structures that neither adapt well to languages with free term order (Covington, 1990), nor look for relationships close to semantic interpretation (Gardent and Kallmeyer, 2003). In contrast, *dependency parsing* captures the relations between a term and its dependents, simplifying the description and extending (Fundel et al., 2007) the use of DG (Tesnière, 1959). However, polynomial time is only achieved in certain cases (Gómez et al., 2009), which suggests that TM should combine information from

both dependencies and constituents, looking for a trade-off between syntactic information and ease of phrase extraction. Here we can take advantage of the lexicalized *tree adjoining grammars* (TAG) (Joshi, 1969), a type of MCSG for which the derivation controller can be interpreted as a dependency graph (Candito and Kahane, 1998), allowing the modelling of a dependency parser from rich constituency information. To give this approach a practical sense it is necessary to reduce the combinatorial explosion of trees associated to lexicalization and extended domain of locality, which can be solved by means of tree factorization (de la Clergerie, 2010).

2.2.2. Semantic modeling

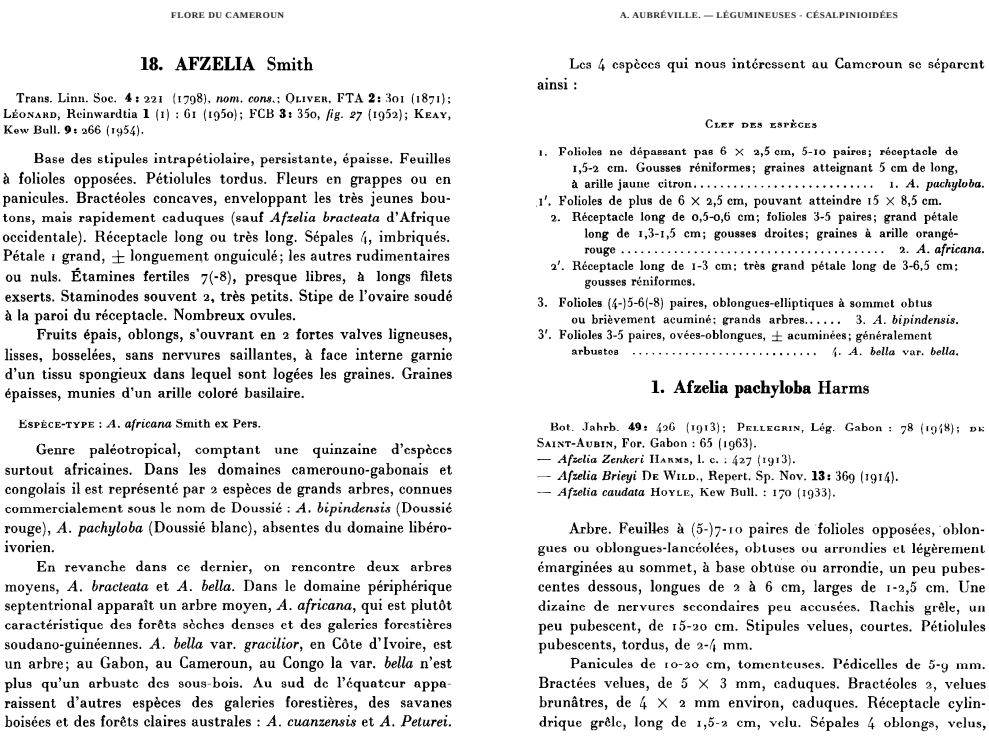
We seek to support searching and reasoning facilities, but at the same time express content in a form that is logically precise, humanly readable and computationally tractable (Sowa, 1984). This takes us away from formalisms such as *region algebras* (Clarke et al., 1995), which require structured texts (Miyao et al., 2006), and leads us to focus on the so called knowledge-based ones: *description logics* (DL) and *network-based systems*. The former (Baader et al., 2003) use a variant of FOL, in which reasoning amounts to verifying logical consequence, which provides a decidable and declarative basis for KD. In network-based proposals, knowledge is represented by means of graph-like structures, and reasoning is accomplished by procedures that manipulate them. We here include *semantic networks* (Richens, 1956) and *frames* (Minsky, 1974), both of which suffer from the absence of a well defined semantics that translates into a lack of declarative power (Björne et al., 2009), including difficulty in handling negation. More recently, *conceptual graphs* (CG) (Sowa, 1976) have the expressing power of FOL. This justifies the consideration of decidable fragments such as the *simple conceptual graphs* (SG), which correspond to existentially quantified conjunctions of atoms. Reasoning is then introduced on the basis of a graph morphism called *projection* (Baget and Mugnier, 2002), which proves to be both sound and complete with regard to deduction.

The graph structure of CG seems to provide a greater expressiveness than the tree one of most DL (Delteil and Faron, 2002), with two substantial differences between both formalisms. The former refers to the incorporation of both a terminological and an assertional language in DL, while CG directly represents knowledge in a graphical way. The second is that DL are characterized by the universally quantified role restriction, which is not present in CG. All of this justifies the interest aroused by CG in the NLP community (Baader et al., 2003), while DL are mostly widely known as the basis of ontology languages in areas such as biology and the semantic web (Horrocks, 2005). Thus, CG seem to be better adapted for TM, but exploiting their properties depends on the ability to access environments in which they can be automatically generated from source documents, something natural to dependency parse relations (Parapatics and Dittenbach, 2009).

2.3. Our contribution

In order to provide full TM capabilities, we describe an NLP-based KE protocol that uses SG as semantic representation. The proposal is organized as a chain of lexical, syntactic and semantic analysis, our contribution focusing on this latter task. Here, we describe a knowledge adquisition process on primary relationships between tokens

identified by a dependency parse built from the output of a POST system and a lexicalixed TAG with a high degree of tree-factorization interpreted in dynamic programming.



FLORE DU CAMEROUN

18. AFZELIA Smith

Trans. Linn. Soc. **4**: 221 (1798), *nom. cons.*; OLIVER, FTA **2**: 301 (1871); LÉONARD, Reinwardtia **1** (1) : 61 (1950); FCB **3**: 350, *fig. 27* (1952); KEAY, Kew Bull. **9**: 266 (1954).

Base des stipules intrapétiolaire, persistante, épaisse. Feuilles à folioles opposées. Pétiolules tordus. Fleurs en grappes ou en panicules. Bractéoles concaves, enveloppant les très jeunes boutons, mais rapidement caduques (sauf *Afzelia bracteata* d'Afrique occidentale). Réceptacle long ou très long. Sépales 4, imbriqués. Pétale 1 grand, ± longuement onguiculé; les autres rudimentaires ou nuls. Étamines fertiles 7(-8), presque libres, à longs filets exserts. Staminodes souvent 2, très petits. Stipe de l'ovaire soudé à la paroi du réceptacle. Nombreux ovules.

Fruits épais, oblongs, s'ouvrant en 2 fortes valves ligneuses, lisses, bosselées, sans nervures saillantes, à face interne garnie d'un tissu spongieux dans lequel sont logées les graines. Graines épaisses, munies d'un arille coloré basilaire.

ESPÈCE-TYPE : *A. africana* Smith ex Pers.

Genre paléotropical, comptant une quinzaine d'espèces surtout africaines. Dans les domaines camerouno-gabonais et congolais il est représenté par 2 espèces de grands arbres, connues commercialement sous le nom de Doussié : *A. bipindensis* (Doussié rouge), *A. pachyloba* (Doussié blanc), absentes du domaine libéro-ivorien.

En revanche dans ce dernier, on rencontre deux arbres moyens, *A. bracteata* et *A. bella*. Dans le domaine périphérique septentrional apparaît un arbre moyen, *A. africana*, qui est plutôt caractéristique des forêts sèches denses et des galeries forestières soudano-guinéennes. *A. bella* var. *gracilior*, en Côte d'Ivoire, est un arbre; au Gabon, au Cameroun, au Congo la var. *bella* n'est plus qu'un arbuste des sous-bois. Au sud de l'équateur apparaissent d'autres espèces des galeries forestières, des savanes boisées et des forêts claires australes : *A. cuanzensis* et *A. Peturei*.

A. AUBRÉVILLE. — LÉGUMINEUSES - CÉSALPINIOIDÉES

Les 4 espèces qui nous intéressent au Cameroun se séparent ainsi :

CLEF DES ESPÈCES

1. Folioles ne dépassant pas 6 × 2,5 cm, 5-10 paires; réceptacle de 1,5-2 cm. Gousses réniformes; graines atteignant 5 cm de long, à arille jaune citron..... 1. *A. pachyloba*.
1′. Folioles de plus de 6 × 2,5 cm, pouvant atteindre 15 × 8,5 cm.
2. Réceptacle long de 0,5-0,6 cm; folioles 3-5 paires; grand pétale long de 1,3-1,5 cm; gousses droites; graines à arille orangé-rouge ..... 2. *A. africana*.
2′. Réceptacle long de 1-3 cm; très grand pétale long de 3-6,5 cm; gousses réniformes.
3. Folioles (4-)5-6(-8) paires, oblongues-elliptiques à sommet obtus ou brièvement acuminé; grands arbres..... 3. *A. bipindensis*.
3′. Folioles 3-5 paires, ovées-oblongues, ± acuminées; généralement arbustes ..... 4. *A. bella* var. *bella*.

1. Afzelia pachyloba Harms

Bot. Jahrb. **49**: 426 (1913); PELLEGRIN, Lég. Gabon : 78 (1948); DE SAINT-AUBIN, For. Gabon : 65 (1963).
— *Afzelia Zenkeri* HARMS, l. c. : 427 (1913).
— *Afzelia Brieyi* DE WILD., Repert. Sp. Nov. **13**: 369 (1914).
— *Afzelia caudata* HOYLE, Kew Bull. : 170 (1933).

Arbre. Feuilles à (5-)7-10 paires de folioles opposées, oblongues ou oblongues-lancéolées, obtuses ou arrondies et légèrement émarginées au sommet, à base obtuse ou arrondie, un peu pubescentes dessous, longues de 2 à 6 cm, larges de 1-2,5 cm. Une dizaine de nervures secondaires peu accusées. Rachis grêle, un peu pubescent, de 15-20 cm. Stipules velues, courtes. Pétiolules pubescents, tordus, de 2-4 mm.

Panicules de 10-20 cm, tomenteuses. Pédicelles de 5-9 mm. Bractées velues, de 5 × 3 mm, caduques. Bractéoles 2, velues brunâtres, de 4 × 2 mm environ, caduques. Réceptacle cylindrique grêle, long de 1,5-2 cm, velu. Sépales 4 oblongs, velus,

Extrait de Aubréville A., 1970. Flore du Cameroun 9. Légumineuses césalpinioidées. Mus. Natl. Hist. Nat., Paris, p. 339.

Figure 1: The description of the genus Afzelia

3. The running corpus

We use as running *corpus* a set of books describing the West African flora: the *"Flore du Cameroun"*, published between 1963 and 2001, produced by different research groups and supplied by the *French Institute of Research for Cooperative Development*. It consists of about forty volumes in French, each one running to about 300 pages. The text is organized taxonomically, introducing genera (resp. species) in separate chapters (resp. sections), and the descriptions include concepts that are related both taxonomically and non-taxonomically. In the first case, they are organized into sub- and super-tree structures, involving the most frequent relationships in biological ontologies: the generic (`"is-a"`), partitive (`"part-of"`) and instance (`"instance-of"`) ones.

Non-taxonomic relations include equivalence and associative links. The first relate to concepts that can be represented by more than one entry, which is not unusual either as the result of error or of the existence of vernacular names in use. The associative case involves thematic links between terms that are neither hierarchical nor equivalent,

but are nevertheless semantically or contextually related to one another. Our reference is the *plant ontology* (PO) database (Jaiswal et al., 2005), including locative relationships ("`adjacent-to`" or "`located-in`") and links representing the functions and processes a concept has or is involved in ("`participates-in`", "`develops-from`", `"derives-by-manipulation-from"` or `"has-participant"`).

Each chapter is organized in sections with a title, a narrative description and a dichotomy, and sections can replicate this structure on subsections. Title includes in its first line the authors, and the taxon family and subfamily we are dealing with. A second line refers to the botanical genus to which the section is devoted, as well as the author who made the discovery. Descriptions relate to morphological aspects such as color, texture, size or form. This implies the presence of nominal sentences, adjectives and also adverbs to express frequency and intensity, and named entities to denote dimensions. A set of keys is included when the range presented has other inferior ones. An example, for the genus *Afzelia*, with a fragment of section is shown in Fig. 1.

Grammatical structures enable us to propagate the relationships through linguistic constructions, as with enumerations on expressions pointing out instances for the color or the form, and the vocabulary is shared by most texts on this matter. We denote this *corpus* by $\mathscr{B}$, its main data set features being a size of 33.9 Gb with 2,719 documents that include a total of 863,297 terms. When it comes to document length, the minimum (resp. maximum) size is 15 (resp. 58,297), the average length being 2,079.46.

4. Simple conceptual graphs and searchable bases

The semantic model is defined with respect to a *support*, which compiles the main concepts, relations and vocabulary that exist in the world we are trying to describe. Most of the definitions are due to (Baget and Mugnier, 2002; Genest and Chein, 2005).

Definition 1. *A* support *is a triple* $\mathcal{S} = (\mathcal{T}_C, \mathcal{T}_\mathcal{R}, \mathcal{I})$ *of finite sets pairwise disjoint, such that* $\mathcal{T}_C$ *(resp.* $\mathcal{T}_\mathcal{R}$*) is a partially ordered set of* concept *(resp.* relation*)* types*. These orders are interpreted as specialization relationships. So,* $t \leq r$ *is read as* r is a generalization of t *or, also, as* r subsumes t*. Types in* $\mathcal{T}_C$ *posses a greater element,* $\top$*, called* universal type*. Types in* $\mathcal{T}_\mathcal{R}$ *may be of any arity greater or equal to 1, and only those with same arity are comparable. The countable set* $\mathcal{I}$ *is a collection of* individual markers *with a* generic marker $* \notin \mathcal{I}$*. The set* $\mathcal{I} \cup \{*\}$ *is partially ordered and its elements pairwise non-comparable, the greatest one being* $*$*.*

We can identify the markers with a dictionary representing lexical forms, while concepts refer to their semantic categories and relations to the relationships between them. Concepts and relations can be linked together in order to describe facts.

Definition 2. *A* simple conceptual graph (SG) defined over a support $\mathcal{S} = (\mathcal{T}_C, \mathcal{T}_\mathcal{R}, \mathcal{I})$ *is a 4-tuple* $\mathcal{G} = (C, \mathcal{R}, \mathcal{E}, \mathcal{L})$*, where* $(C \cup \mathcal{R}, \mathcal{E})$ *is a bipartite multigraph with* C *and* $\mathcal{R}$ *disjoint sets of* concept *and* relation nodes*, respectively.* $\mathcal{E}$ *is the multiset of* edges *and* $\mathcal{L}$ *is a labeling function for nodes and edges. A node* $c \in C$ *is labeled by a pair* $[\text{type(c)}, \text{marker(c)}] \in \mathcal{T}_C \times (\mathcal{I} \cup \{*\})$*. A node* $r \in \mathcal{R}$ *is labeled by* $\text{type}(r) \in \mathcal{T}_\mathcal{R}$ *and the degree of* r*, i.e., the number of edges incident to, must be equal to the arity of* $\text{type}(r)$*. An edge in* $\mathcal{E}$*, labeled by* $i \in \mathbb{N}$*, connecting nodes* $r \in \mathcal{R}$ *and* $c \in C$*, is denoted by* (r, i, c)*.*

The edges $(r, 1, c_1), \ldots, (r, k, c_k)$ *incident to* $r \in \mathcal{R}$ *are totally ordered and labeled from* 1 *to the degree* k *of* r*. We then shortly denote* $r = \text{type}(r)(c_1, \ldots, c_k)$*.*

A SG provides an ontology of the domain, where concepts refer to the markers in the support associating a conceptual type. Reasoning is introduced through subsumption.

Definition 3. *Let* $\mathcal{G}_1 = (C_1, \mathcal{R}_1, \mathcal{E}_1, \mathcal{L}_1)$ *and* $\mathcal{G}_2 = (C_2, \mathcal{R}_2, \mathcal{E}_2, \mathcal{L}_2)$ *be* SG *defined on a support* $\mathcal{S} = (\mathcal{T}_C, \mathcal{T}_\mathcal{R}, \mathcal{I})$*, then a* projection *from* $\mathcal{G}_1$ *to* $\mathcal{G}_2$ *is a mapping* π *from* C_1 *to* C_2*, and from* $\mathcal{R}_1$ *to* $\mathcal{R}_2$ *verifying:*

$$(r, i, c) \in \mathcal{E}_1 \Rightarrow (\pi(r), i, \pi(c)) \in \mathcal{E}_2 \quad \text{and} \quad x \in C_1 \cup \mathcal{R}_1 \Rightarrow \mathcal{L}_2(\pi(x)) \leq \mathcal{L}_1(x) \qquad (1)$$

where, if $x \in C_1$*,* $\leq$ *refers to the cartesian product of the order on* $\mathcal{T}_C$ *and on* $\mathcal{I} \cup \{*\}$[1]*. If* $x \in \mathcal{R}_1$*, then* $\leq$ *refers to the order on* $\mathcal{T}_\mathcal{R}$*. We say that* $\mathcal{G}_1$ subsumes $\mathcal{G}_2$ *or that* $\mathcal{G}_1$ *is* more general than $\mathcal{G}_2$*. The set of projections from* $\mathcal{G}_1$ *to* $\mathcal{G}_2$ *is denoted by* $\text{proj}(\mathcal{G}_1, \mathcal{G}_2)$*.*

A projection from $\mathcal{G}_1$ to $\mathcal{G}_2$ means that the knowledge represented by the first is contained in the one represented by the second, which defines a reasoning model that is logically sound and complete with regard to deduction in FOL and locates the query answering problem in a decidable framework. However, when information needs do not exactly correspond to a projection, we must relax the structural constraints.

Definition 4. *Let* $\mathcal{D}$*,* $\mathcal{D}'$ *and* $\mathcal{Q}$ *be* SG *defined on a support* $\mathcal{S}$*, and* ς *a mapping from the set of* SG *defined on* $\mathcal{S}$ *onto itself, such that* $\varsigma(\mathcal{D}) = \mathcal{D}'$*. If* $\pi \in \text{proj}(\mathcal{Q}, \mathcal{D}')$*, then* (π, ς) *is a* projection from $\mathcal{Q}$ to $\mathcal{D}$ modulo ς*.*

The idea is to supply a set of transformations in order to determine the relevance of a document $\mathcal{D}$ to a query $\mathcal{Q}$, when there is some kind of relation between them.

Definition 5. *Let* $\mathcal{G} = (C, \mathcal{R}, \mathcal{E}, \mathcal{L})$ *be a* SG *defined on a support* $\mathcal{S} = (\mathcal{T}_C, \mathcal{T}_\mathcal{R}, \mathcal{I})$*,* compatible/2 *a binary predicate and* $(t, t') \in (C \times (\mathcal{T}_C \times (\mathcal{I} \cup \{*\}))) \cup (\mathcal{R} \times \mathcal{T}_\mathcal{R})$ *compatible nodes. We define the* substitution of t by t' on $\mathcal{G}$ *as the* SG *obtained replacing* t *by* t'*.*

The result of the join of $c, c' \in \mathcal{T}_C$*, such that* $\mathcal{L}(c) = \mathcal{L}(c')$*, is the* SG *obtained from* $\mathcal{G}$ *by identification of* c *and* c'*. Finally,* adding a node $n \in C \cup \mathcal{R}$*, results on the* SG $\mathcal{G} + \mathcal{N}$*, where* $\mathcal{N}$ *is reduced to* n*. If* $n \in \mathcal{R}$*, neighbors must be specified.*

Compatibility is not necessarily symmetric and is often defined on the basis of some distance between types. As a join can substantially change the structure of an SG, it is is considered more distancing than substitutions. Given that an addition introduces an external element, it is taken to be more complex than a join. The combination of transformations results in four kinds of answer to a given query.

Definition 6. *Let* $\mathcal{D}$ *and* $\mathcal{Q}$ *be* SG *defined on a support* $\mathcal{S}$*. Then* $\mathcal{D}$ *is an* exact answer to $\mathcal{Q}$ *iff* $\text{proj}(\mathcal{Q}, \mathcal{D}) \neq \emptyset$*. It is an* approximate answer to $\mathcal{Q}$ *when there exists a sequence of substitutions* ς*, such that* $\text{proj}(\mathcal{Q}, \varsigma(\mathcal{D})) \neq \emptyset$*.*

[1] i.e., $(type(\pi(x)), marker(\pi(x))) \leq (type(x), marker(x))$ iff $type(\pi(x)) \leq type(x)$, and $marker(\pi(x)) \leq marker(x)$.

As exact answers are a rare case of approximate ones, we use this last term to refer to both categories. In order to further increase the degree of flexibility associated to querying, we can also include joins and node adds as admissible transformations.

Definition 7. *Let $\mathcal{D}$ be an* SG *defined on a support $\mathcal{S}$. We say that a* sequence ς of substitutions and joins *(resp.* and node adds*)* is acceptable *iff ς does not contain too many joins (resp. node adjunctions) relative to the number of nodes in $\mathcal{D}$ (resp. and ς* is acceptable for the joins*). The ratio numbers of joins (μ_j) and node adds (μ_a) can be chosen by the user.*

Definition 8. *Let $\mathcal{D}$ and $\mathcal{Q}$ be* SG *defined on a support $\mathcal{S}$. We say that $\mathcal{D}$ is a* plausible *(resp.* partial*)* answer *to $\mathcal{Q}$ iff there is an acceptable sequence ς of substitutions and joins (resp. and node adds), such that* $\text{proj}(\mathcal{Q}, \varsigma(\mathcal{D})) \neq \emptyset$.

We now introduce, from a partial order in the set of transformations, a ranking protocol to show the user the answers in descending order of relevance.

Definition 9. *Given a support $\mathcal{S}$, let $\mathcal{Q}$ and $\mathcal{D} = \{\mathcal{D}_i\}_{i\in I}$ be the* SG *associated to a query and a document database, and let $\mathcal{A}_{\mathcal{Q}}^{\mathcal{D}}$ be the collection of answers obtained through a set $\mathcal{T}_{\mathcal{Q}}^{\mathcal{D}}$ of graph transformation sequences applied to get a projection of $\mathcal{Q}$ on some $\mathcal{D}_i$, $i \in I$. We then define a* ranking function associated to $\mathcal{Q}$ and $\mathcal{D}$ *as the ordering naturally induced in $\mathcal{A}_{\mathcal{Q}}^{\mathcal{D}}$ by any partial order on $\mathcal{T}_{\mathcal{Q}}^{\mathcal{D}}$.*

We consider an approximate (resp. plausible) answer more relevant than a plausible (resp. partial) one. For a same type, relevance is inversely proportional to the number of transformations applied. No explicit document length normalization (resp. graph-based term weighting) is applied, since we assume the scale is provided by graph-ranking computation (resp. the results seem to be similar, despite its simplicity).

Definition 10. *Given a support $\mathcal{S}$, let $\mathcal{Q}$ and $\mathcal{D} = \{\mathcal{D}_i\}_{i\in I}$ be the* SG *associated to a query and a document database, and let $\mathcal{A}_{\mathcal{Q}}^{\mathcal{D}}$ be the collection of answers obtained through a set $\mathcal{T}_{\mathcal{Q}}^{\mathcal{D}}$ of graph transformation sequences applied on $\mathcal{Q}$ to get a projection on some $\mathcal{D}_i$, $i \in I$. We define the* Genest's partial order on $\mathcal{T}_{\mathcal{Q}}^{\mathcal{D}}$ *as:*

$$t <_G t' \text{ iff } \begin{cases} t' & \text{associates approximate answer OR} \\ t & \text{associates a partial answer OR} \\ t \text{ (resp. } t') & \text{associates a partial (resp. plausible) answer OR} \\ t,\ t' & \text{associate the same type of answer AND } \mid t \mid > \mid t' \mid \end{cases}$$

while that $t =_G t'$ iff t AND t' associate the same type of answer AND $\mid t \mid = \mid t' \mid$

5. Knowledge extraction

For this purpose, the kernel of our contribution, we contemplate a chain of lexical, syntactic and semantic analysis that performs TM with minimal user intervention.

5.1. The lexical frame

We do not require specific POST systems. The only condition provided is on the output, which must include all possible lexical categories for a given occurrence of a form and is denoted as indicated below, introducing some additional structural details in order to later integrate semantic data. In practice, we chose the *Alexina* architecture (Sagot, 2010), which is based on a finite state morphology that combines its output with lexical information retrieved from a lexicon for French called LEFFF.

Definition 11. *Let* $\{s_i\}_{1\leq i\leq n}$ *be the sequence of sentences in a* corpus $\mathscr{C}$ *and* $\Theta_{i,j},\ 1 \leq j \leq |\ s_i\ |$ *be the* occurrence *of a* form *in the i-th sentence,* s_i*. We denote the association of the lexical category (a) and semantic class (b) to this form, in this sentence, by* $\Theta^{a,b}_{i,j}$ *and we call it* term*. We use an anonymous-variable notation,* $\Theta^{a,\_}_{i,j}$*, in order to designate the set of terms that can only be differentiated by their semantic class, which we call* token*. We denote by* $\Theta^{\_,\_}_{i,j}$ *the set of tokens referring to the same occurrence of a form, which we call* cluster*.*

We also consider a free-variable notation, using capital letters, in order to enumerate a range of values. So, for example, $\Theta^{a,X}_{i,j}$ *refers to the sequence of terms in the token* $\Theta^{a,\_}_{i,j}$*, whose semantic class X is applicable in that context. We can naturally extend this notation to occurrences of tokens and clusters.*

We illustrate the notation in Fig. 2 for the sentence *"feuilles à nervures denticulées"* (`"leaves with veins dentate"`). Terms are represented by triangles, tokens by ellipses and clusters by rectangles. The semantic classes are taken from Table 3.

5.2. The parsing frame

The proposal does not depend of any particular frame, although for the reasons outlined above our choice fell on a TAG for French, applying a high degree of abstraction in dynamic programming and using meta-grammars (de la Clergerie, 2010). The parse graphically compiles the head-dependent relationships within the text analyzed, as shown in Fig. 3 by dotted lines connecting the nodes involved in each case. We can observe the impact that both lexical and syntactic ambiguities have on the number of possible dependencies that go forward to the semantic analysis stage. In the first case, they multiply in relation to the number of tokens in a single cluster. In the second, we can see an analogue effect resulting from the multiplication of dependencies on the modifiers. An example of this would be *"denticulées"* (`"dentate"`), which could be a modifier of either *"feuilles"* (`"leaves"`) or *"nervures"* (`"veins"`) in Fig. 3. This is a well-known phenomenon linked to the association of prepositional attachments to a nominal phrase, and which here provides us with two possible interpretations for the sentence: `"leaves with dentate veins"` or, alternatively, `"dentate leaves with veins"`.

There are still situations in which ambiguities are exclusively of semantic origin. An example is the use of coordination structures relating entities to a list of adjectives, as in *"des sépales ovales-aigus, glabres ou éparsement hérissés"* (`"Sepals oval-pointed, smooth or scattered bristly"`), where the property *"hérissés"* (`"bristly"`) could be attached to the adjectives *"glabres"* (`"smooth"`) or *"ovales-aigus"* (`"oval-pointed"`). Here, the only way to solve the problem is to understand

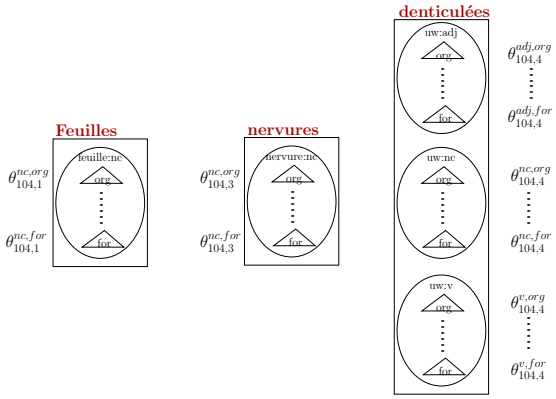


Figure 2: Lexical notation

the precise nature of the plant organs concerned. Since an ambiguity corresponds to a situation where a dependent token has more than one head token, solving it results in filtering out the less plausible dependencies in favor of the most plausible ones.

5.3. The semantic frame

We now prioritize the dependencies, gathering data from the text in order to extract its meaning. We consider three steps, the first two of which are aimed at exploiting the sequence of structures obtained from the previous lexical and syntactic analysis stages, classifying any ambiguities according to their order of priority. The third one determines the semantic information is involved in each link. These steps extrapolate the estimations from a local level (sentence) to a global one (*corpus*) or, in other words, initial data obtained at sentence level are combined and evaluated throughout the whole *corpus* in order to extract new conclusions that can then be applied in each sentence, the process recommencing iteratively. This is needed to extend some previous notation.

Definition 12. *Let* $\{s_i\}_{1\leq i\leq n}$ *be the sequence of sentences in a* corpus $\mathscr{C}$ *and* $\Theta_{i,j}$, $1 \leq j \leq |\ s_i\ |$ *be the* occurrence *of a form in the i-th sentence,* s_i*. We denote the association of the lexical category (a) and semantic class (b) to this form, anywhere in* $\mathscr{C}$*, by* $\tilde{\Theta}^{a,b}_{i,j}$ *and we call it* plausible term*. We also naturally extend the anonymous-variable (resp. free-variable) notation previously introduced for terms, tokens and clusters.*

We also need some notation for managing head-dependent relationships at sentence (resp. *corpus*) level. We have to refer to transitions between tokens (resp. plausible tokens) that constitute the parser output and to the sets of transitions between tokens from two different clusters (resp. plausible clusters). Finally, we have to deal with transitions between terms (resp. plausible terms) for semantic categorization.

Definition 13. *Let* s_i, $1 \leq i \leq n$ *be the i-th sentence in a* corpus $\mathscr{C}$ *and* τ *be the sequence of the grammar rules necessary to generate the token* $\Theta^{c,\_}_{i,k}$ *from the token* $\Theta^{a,\_}_{i,j}$ *in the head-dependent graph. We denote the* dependency between $\Theta^{a,\_}_{i,j}$ and $\Theta^{c,\_}_{i,k}$,

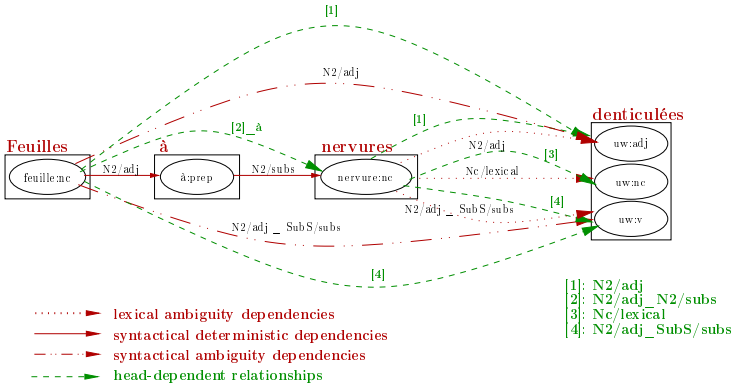


Figure 3: Head-dependent relationships

labeled by τ*, as* $\delta^{\theta^{d,-}_{i,j},\tau,\theta^{c,-}_{i,k}}$*. The notation can be naturally extended to terms, clusters and plausible structures, and we talk then about* plausible dependencies*.*

5.3.1. Categorization of tokens

The goal is to compute which, for each cluster, is the most probable token. Namely, we want to determine the category for each occurrence of a given form in a sentence. The process corresponds to the equations in Table 1, which we comment on below:

(2). We start by calculating the local probability, at sentence level, that can be associated to a token in a cluster. This is a ratio that depends on the number of tokens involved in the said cluster. If there is only one, its probability is 1.

(3). This defines the global probability of a plausible token in the *corpus*, at iteration $n+1$. It is a proportion of the local probability associated with tokens of the same category and form as those of the token in question, in relation to the probability when the category is free.

(4). It determines the value of the local probability that can be associated with a token in a cluster, at iteration $n+1$. In order to do so, we allocate the probabilities calculated globally, distributing them proportionally between the global probabilities of the plausible tokens associated with the cluster.

The iterations continue until convergence at a fixed point, or until a fixed approximation threshold υ_{to} is achieved on local probabilities. Alternatively or simultaneously, we can fix a maximum number of iterations ι_{to} to apply.

5.3.2. Categorization of dependencies between tokens

The objective is to measure the viability of the syntactic dependencies generated by the parser between the previously categorized tokens. We once again opt for an iterative strategy, here determined by the equations in Table 2, which we now describe:

(5). An initial weight is first associated to each syntactic dependency depending on its label. We thereby seek to assign more importance to those shared by a greater number of parses, amongst those sharing a single dependent cluster.

$$P(\Theta^{a,\text{-}}_{i,j})_{\text{local(0)}} = \frac{1}{|\{\Theta^{X,\text{-}}_{i,j}\}|} \tag{2}$$

$$P(\tilde{\Theta}^{a,\text{-}}_{i,j})_{\text{global(n+1)}} = \frac{\sum\limits_{\Theta_{k,l}=\Theta_{i,j}} P(\Theta^{a,\text{-}}_{k,l})_{\text{local(n)}}}{\sum\limits_{\Theta^{X,\text{-}}_{k,l},\Theta_{k,l}=\Theta_{i,j}} P(\Theta^{X,\text{-}}_{k,l})_{\text{local(n)}}} \tag{3}$$

$$P(\Theta^{a,\text{-}}_{i,j})_{\text{local(n+1)}} = \frac{P(\tilde{\Theta}^{a,\text{-}}_{i,j})_{\text{global(n+1)}}}{\sum\limits_{\Theta^{X,\text{-}}_{k,l},\Theta_{k,l}=\Theta_{i,j}} P(\tilde{\Theta}^{X,\text{-}}_{k,l})_{\text{global(n+1)}}} \tag{4}$$

Table 1: Model for categorization of tokens

(6). We calculate the local probabilities for the syntactic dependencies. Given that these are characterized by their head and dependent tokens, and by their label, we make these probabilities depend on the local ones of such tokens, as well as on the weight assigned to the associated label. This is calculated as a proportion of the above-mentioned values for the syntactic dependency in question, in relation to the set of dependencies associated with the dependent token cluster.

(7). This defines the global probability of a plausible dependency at iteration $n + 1$. It is calculated as a proportion of the local one associated with syntactic dependencies coinciding with the one under consideration (except in the locating sentence), in relation to the set of local dependencies associated with dependent tokens that also coincide with the one under consideration (except in the locating cluster).

(8). This establishes the value of the local probability of a dependency in iteration $n + 1$. To this end we allocate the probabilities calculated globally, distributing them proportionally amongst the global probabilities of the plausible syntactic dependencies associated with the dependent tokens coinciding with the one under consideration (except in the locating cluster).

The process repeats itself until it converges at a fixed point, or until a fixed approximation threshold υ_{dto} is achieved on local probabilities. Also, alternatively or simultaneously, we can fix a maximum number of iterations ι_{dto} to apply.

5.3.3. Categorization of dependencies between terms

The goal is to attach the semantic classes to the tokens involved inone and the same syntactic dependency, in order to identify the semantic ones between terms in two clusters. Thus, given a dependent term, we seek to define its head by means of the syntactic dependencies categorized previously. We first need to introduce some notation.

Definition 14. *Let s_i, $1 \leq i \leq n$ be the i-th sentence in a* corpus *$\mathscr{C}$, and $\mathscr{T}$ (resp. $\mathscr{F}$) be the set of semantic classes (resp. forms) associated to $\mathscr{C}$ (resp. to $\mathscr{T}$) by means*

$$W(\delta^{\Theta^{a,-}_{i,j},\tau,\Theta^{b,-}_{i,k}}) = \frac{|S \overset{*}{\Rightarrow} \Theta^{a,-}_{i,j} \overset{\tau}{\Rightarrow} \Theta^{b,-}_{i,k}|}{\sum_{\delta^{\Theta^{Y,-}_{i,X},T,\Theta^{Z,-}_{i,k}}} |S \overset{*}{\Rightarrow} \Theta^{Y,-}_{i,X} \overset{T}{\Rightarrow} \Theta^{Z,-}_{i,k}|} \tag{5}$$

$$P(\delta^{\Theta^{a,-}_{i,j},\tau,\Theta^{b,-}_{i,k}})_{\text{local(0)}} = \frac{P(\Theta^{a,-}_{i,j})_{\text{local}} \cdot P(\Theta^{b,-}_{i,k})_{\text{local}} \cdot W(\delta^{\Theta^{a,-}_{i,j},\tau,\Theta^{b,-}_{i,k}})}{\sum_{\Theta^{Y,-}_{i,X},\Theta^{Z,-}_{i,k},\delta^{\Theta^{Y,-}_{i,X},T,\Theta^{Z,-}_{i,k}}} P(\Theta^{Y,-}_{i,X})_{\text{local}} \cdot P(\Theta^{Z,-}_{i,k})_{\text{local}} \cdot W(\delta^{\Theta^{Y,-}_{i,X},T,\Theta^{Z,-}_{i,k}})} \tag{6}$$

$$P(\delta^{\tilde{\Theta}^{a,-}_{i,j},\tau,\tilde{\Theta}^{b,-}_{i,k}})_{\text{global(n+1)}} = \frac{\sum_{\Theta_{l,m}=\Theta_{i,j},\Theta_{l,p}=\Theta_{i,k}} P(\delta^{\Theta^{a,-}_{l,m},\tau,\Theta^{b,-}_{l,p}})_{\text{local(n)}}}{\sum_{\delta^{\Theta^{Y,-}_{l,X},T,\Theta^{Z,-}_{l,p}},\Theta_{l,p}=\Theta_{i,k}} P(\delta^{\Theta^{Y,-}_{l,X},T,\Theta^{Z,-}_{l,p}})_{\text{local(n)}}} \tag{7}$$

$$P(\delta^{\Theta^{a,-}_{i,j},\tau,\Theta^{b,-}_{i,k}})_{\text{local(n+1)}} = \frac{P(\delta^{\tilde{\Theta}^{a,-}_{i,j},\tau,\tilde{\Theta}^{b,-}_{i,k}})_{\text{global(n+1)}}}{\sum_{\delta^{\tilde{\Theta}^{Y,-}_{l,X},T,\tilde{\Theta}^{Z,-}_{l,m}},\Theta_{l,m}=\Theta_{i,k}} P(\delta^{\tilde{\Theta}^{Y,-}_{l,X},T,\tilde{\Theta}^{Z,-}_{l,m}})_{\text{global(n+1)}}} \tag{8}$$

Table 2: Model for categorization of dependencies between tokens

of some reliable technique. We then denote by $\mathscr{F}(b)$ the subset of forms associated to $b \in \mathscr{T}$, and we say that $\Theta^{a,b}_{i,j}$, $1 \leq j \leq |\ s_i\ |$ is a stable term *iff $b \in \mathscr{T}$ and $\Theta_{i,j} \in \mathscr{F}(b)$.*

Entities	**Lemmas** (in French)
organ	fleur, staminode, tige, feuille, hypanthe, périanthe, rameau, ...
fruit	fruit, samare, drupe, capsule, akène, ...
Properties	**Lemmas** (in French)
color	verdâtre, violacé, noirâtre, violet, jaunâtre, orange, roux, rose, ...
form	obconique, oblancéolé, oblong, bifolié, crateriforme, punctiforme, ...
size	moyen, petit, double, épais, inégal, entier, longue, ...
texture	hispide, bifide, globuleux, coriace, velutineux, gélatineux, barbu, ...
position	antérieur, dessus, voisin, seul, latéral, transversal, ...

Table 3: The set $\mathscr{T}$ of initial semantic classes (types) for the *corpus* $\mathscr{B}$

Intuitively, a term is stable when we have reliable information about the correspondence between its semantic class and its form, obtained either from the user or by means of a method held to be completely trustworthy. Our proposal contemplates the use of both mechanisms. On the one hand, the user defines the set of semantic categories that in our running *corpus* $\mathscr{B}$ are organized as entities ($\mathscr{E}$) and properties ($\mathscr{P}$), together with a set of initial associated forms such as the one shown in Table 3. On the other, the system makes use of *collocations*, sequences of words that co-occur more often than would be expected by chance and in which they keep their original meaning,

in contrast to the case of *locutions*. The idea is to filter out the parse in order to locate collocations that enable a form to be associated with a semantic class.

Word (in French)	**Position**	**Class**	**Word** (in French)	**Position**	**Class**
teinté	[2]	color	épaisseur	[1]	size
texture	[2]	texture	atteindre	[1]	organ/fruit

Table 4: A sample section from the collocations file for *corpus* $\mathscr{B}$

We represent a collocation as a triple of the form *marker-position-semantic class*. The marker serves to identify the collocation for which the form in the indicated position is associated with the class, as shown in Table 4 for the *corpus* $\mathscr{B}$. So, in the sentence *"teintées de rose"* (`"rose-tinted"`), the presence of the marker *"teinté"* (`"tinted"`) reveals that *"rose"* (`"rose"`) is an instance of the class `"color"`. The process thus corresponds to the equations in Table 5, which we now describe:

(9). Before commencing, we give each token a weight verifying the condition presented, the value of which we justify below.

(10). We distribute the weight calculated from Equation 9 evenly between the stable terms. So, the weight we associate with non-stable terms in this token is lower than that associated with the former, giving initial preference to the stable terms.

(11). Iterations commence with the calculation of the local probabilities for semantic dependencies. Since the latter are characterized by their head and dependent terms, together with the syntactic dependency between their associated tokens, we make this value depend on the weights associated with the said terms, as well as on the local probability corresponding to the syntactic dependency. This is calculated as a proportion of the said values for the semantic dependency in question, in relation to the set of dependencies associated with the dependent term cluster.

(12). We define the global probability of a plausible semantic dependency at iteration $n+1$. It is calculated as a proportion of the local probability associated with the semantic dependencies that coincide with the one under consideration (except in the locating sentence), in relation to the set of the local ones associated with the dependent terms that also coincide with the one under consideration (except in the locating cluster).

(13). This establishes the value of the local probability to be associated with a semantic dependency at iteration $n+1$. We allocate the globally calculated probabilities, distributing them proportionally between the global ones of the plausible semantic dependencies associated with dependent terms that coincide with the one under consideration (except in the locating cluster).

The process repeats itself until it converges at a fixed point, or until a fixed approximation threshold υ_{dte} is achieved on local probabilities. Also, alternatively or simultaneously, we can fix a maximum number of iterations ι_{dte} to apply. We call the resulting structure the *semantic of the corpus* $\mathscr{C}$ we are working with.

$$W(\Theta^{a,\text{-}}_{i,j}) > \frac{|\{\Theta^{a,X}_{i,j}\}_{X\in\mathscr{T},\Theta_{i,j}\in\mathscr{F}(X)}|}{|\{\Theta^{a,X}_{i,j}\}_{X\in\mathscr{T}}|} \subseteq (0,1] \tag{9}$$

$$W(\Theta^{a,b}_{i,j}) = \frac{W(\Theta^{a,\text{-}}_{i,j})}{|\{\Theta^{a,X}_{i,j}\}_{X\in\mathscr{T},\Theta_{i,j}\in\mathscr{F}(X)}|} \text{ if } \Theta_{i,j}\in\mathscr{F}(b), \text{ and } \frac{1-W(\Theta^{a,\text{-}}_{i,j})}{|\{\Theta^{a,X}_{i,j}\}_{X\in\mathscr{T},\Theta_{i,j}\notin\mathscr{F}(X)}|} \text{ otherwise} \tag{10}$$

$$P(\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}})_{\text{local}(0)} = \frac{P(\delta^{\Theta^{a,\text{-}}_{i,j},\tau,\Theta^{c,\text{-}}_{i,k}})_{\text{local}}\cdot W(\Theta^{a,b}_{i,j})\cdot W(\Theta^{c,d}_{i,k})}{\sum_{\Theta^{Y,Z}_{i,X},\Theta^{V,W}_{i,k},\delta^{\Theta^{Y,\text{-}}_{i,X},T,\Theta^{V,\text{-}}_{i,k}}} P(\delta^{\Theta^{Y,\text{-}}_{i,X},T,\Theta^{V,\text{-}}_{i,k}})_{\text{local}}\cdot W(\Theta^{Y,Z}_{i,X})\cdot W(\Theta^{V,W}_{i,k})} \tag{11}$$

$$P(\delta^{\tilde{\Theta}^{a,b}_{i,j},\tau,\tilde{\Theta}^{c,d}_{i,k}})_{\text{global}(n+1)} = \frac{\sum_{\Theta_{l,m}=\Theta_{i,j},\Theta_{l,p}=\Theta_{i,k}} P(\delta^{\Theta^{a,b}_{l,m},\tau,\Theta^{c,d}_{l,p}})_{\text{local}(n)}}{\sum_{\delta^{\Theta^{Y,Z}_{l,X},T,\Theta^{V,W}_{l,p}},\Theta_{l,p}=\Theta_{i,k}} P(\delta^{\Theta^{Y,Z}_{l,X},T,\Theta^{V,W}_{l,p}})_{\text{local}(n)}} \tag{12}$$

$$P(\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}})_{\text{local}(n+1)} = \frac{P(\delta^{\tilde{\Theta}^{a,b}_{i,j},\tau,\tilde{\Theta}^{c,d}_{i,k}})_{\text{global}(n+1)}}{\sum_{\delta^{\tilde{\Theta}^{Y,Z}_{l,X},T,\tilde{\Theta}^{V,W}_{l,m}},\Theta_{l,m}=\Theta_{i,k}} P(\delta^{\tilde{\Theta}^{Y,Z}_{l,X},T,\tilde{\Theta}^{V,W}_{l,m}})_{\text{global}(n+1)}} \tag{13}$$

Table 5: Model for categorization of dependencies between terms

Definition 15. *Let $\{s_i\}_{1\le i\le n}$ be the sequence of sentences in a* corpus *$\mathscr{C}$, and $\mathscr{T}$ (resp. $\mathscr{F}$) be the set of semantic classes (resp. forms) associated to $\mathscr{C}$ (resp. to $\mathscr{T}$) by means of some reliable technique. We then define the* semantic of the corpus *$\mathscr{C}$ as:*

$$\mathscr{S}_{\mathscr{C}} := \{\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}},\ P(\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}})_{local} = \max\{P(\delta^{\Theta^{X,Y}_{i,j},Z,\Theta^{V,W}_{i,k}})_{local}\}\} \tag{14}$$

where max *is the maximal function on $\mathbb{N}$, and $\delta^{\Theta^{X,Y}_{i,j},Z,\Theta^{V,W}_{i,k}}$ are the dependencies computed as result of the process previously described. We can restrict the concept to refer the* semantic of a document *$\mathscr{D}$ (resp.* of a sentence *s_i) in $\mathscr{C}$ by*

$$\mathscr{S}^{\mathscr{D}}_{\mathscr{C}} := \{\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}} \in \mathscr{S}_{\mathscr{C}},\ s_i\in\mathscr{D}\} \quad (\textit{resp. by } \mathscr{S}^{\mathscr{D},s_i}_{\mathscr{C}} := \{\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}} \in \mathscr{S}^{\mathscr{D}}_{\mathscr{C}}\}) \tag{15}$$

The semantic of a text is the set of most probable dependencies between its terms, serving as basis for the knowledge representation. We illustrate in Fig. 4, the result of the process for the graph in Fig. 3, which highlights the simplifications made.

5.4. Knowledge representation

We use SG as semantic frame so, although the proposal is independent of the domain knowledge, we need to locate the work in a specific one, in order to suitably model the support. As the choice fell upon the *corpus $\mathscr{B}$*, we retake the set $\mathscr{T}$ of semantic classes (types) shown in Table 3, in order to introduce a partial order on it as follows:

$$\forall\ t\in\mathscr{E} = \{\textit{fruit, organe}\},\ t\le\varepsilon\le\top \tag{16}$$

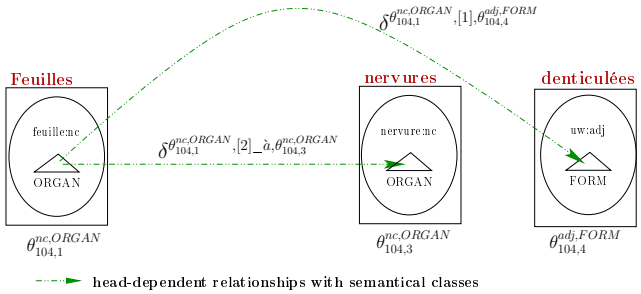


Figure 4: The semantic of a sentence

$$\forall\ t \in \mathscr{P} = \{\textit{couleur, forme, taille, texture, position}\},\ t \leq \rho \leq \top \tag{17}$$

where ε (resp. ρ) is the greater element for the entities (resp. properties) $\mathscr{E}$ (resp. $\mathscr{P}$). In this way, we introduce our running support $\mathcal{S}_{\mathscr{B}} = (\mathcal{T}_{C\mathscr{B}}, \mathcal{T}_{\mathcal{R}\mathscr{B}}, \mathcal{I}_{\mathscr{B}})$ by defining:

$$\mathcal{T}_{C_{\mathscr{B}}} := \{\varepsilon, \rho\} \cup \mathscr{E} \cup \mathscr{P} \cup \{\top\} \tag{18}$$

$$\mathcal{T}_{\mathcal{R}_{\mathscr{B}}} := \{[b,\tau,d], [b,*,d],\ \exists\ \delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}} \in \mathscr{S}_{\mathscr{B}}\} \cup \{[\varepsilon,*,\varepsilon]\} \cup \{[\varepsilon,*,\rho]\} \cup \{[\rho,*,\rho]\} \cup \{[\top,*,\top]\} \tag{19}$$

$$\mathcal{I}_{\mathscr{B}} := \{\Theta^{a,\_}_{i,j},\ \Theta^{c,\_}_{i,k}\}_{\delta^{\Theta^{a,\_}_{i,j},\_,\Theta^{c,\_}_{i,k}}} \tag{20}$$

The relations in $\mathcal{T}_{\mathcal{R}_{\mathscr{B}}}$ summarize transitions between two terms from the point of view of the semantic classes (types) involved. We also add triples representing any transition between the semantically related generic concepts. The partial order in $\mathcal{T}_{C_{\mathscr{B}}}$ and $\mathcal{T}_{\mathcal{R}_{\mathscr{B}}}$ is induced by the one defined in $\mathscr{T}$, and the markers $\mathcal{I}_{\mathscr{B}}$ are defined as the set of forms in $\mathscr{B}$. In this context, we introduce SG on this support from the semantic $\mathscr{S}_{\mathscr{D}_m}$ associated with each of the M documents in the *corpus* $\mathscr{B} = \bigcup_{m \in M} \mathscr{D}_m$, as follows:

$$\mathcal{C}_{\mathscr{D}_m} := \{\Theta^{a,b}_{i,j}, \Theta^{c,d}_{i,k}\}_{\delta^{\Theta^{a,b}_{i,j},\_,\Theta^{c,d}_{i,k}} \in \mathscr{S}_{\mathscr{D}_m}} \qquad \mathcal{R}_{\mathscr{D}_m} := \{[b,\tau,d],\ \exists\ \delta^{\Theta^{\_,b}_{\_,\_},\tau,\Theta^{\_,d}_{\_,\_}} \in \mathscr{S}_{\mathscr{D}_m}\} \tag{21}$$

$$\mathcal{E}_{\mathscr{D}_m} := \bigcup_{\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}} \in \mathscr{S}_{\mathscr{D}_m}} \{([b,\tau,d], 1, \Theta^{a,b}_{i,j}), ([b,\tau,d], 2, \Theta^{c,d}_{i,k})\} \tag{22}$$

$$\mathcal{L}_{\mathscr{D}_m}(X) := \begin{cases} [b, \Theta^{a,\_}_{i,j}] & \text{if } X = \Theta^{a,b}_{i,j} \in C_{\mathscr{D}_m}, \quad 1 \quad \text{if } X = (\_, 1, \_) \in \mathcal{E}_{\mathscr{D}_m} \\ X & \text{if } X \in \mathcal{R}_{\mathscr{D}_m}, \qquad\quad 2 \quad \text{if } X = (\_, 2, \_) \in \mathcal{E}_{\mathscr{D}_m} \end{cases} \tag{23}$$

Succintly, a conceptual node in $C_{\mathscr{D}_m}$ is any term involved in the semantic $\mathscr{S}_{\mathscr{D}_m}$, while relation nodes in $\mathcal{R}_{\mathscr{D}_m}$ are elements of $\mathcal{T}_{\mathcal{R}_{\mathscr{B}}}$ associated to transitions in $\mathscr{S}_{\mathscr{D}_m}$. The multiset of edges $\mathcal{E}_{\mathscr{D}_m}$ contains only binary relations, the head (resp. dependent) term corresponding to the first (resp. second) triple. With regard to $\mathcal{L}_{\mathscr{D}_m}$, this makes it possible to recover the class and the token associated to a given term representing a concept, whilst implementing the identity on the relations, since in our case we build these directly from $\mathscr{S}_{\mathscr{D}_m}$. Its value on edges identifies head (1) and dependent edges

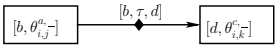


Figure 5: A representation as SG for a dependency $\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}}$

(2). We define the compatibility for nodes $(t, t') \in (\mathcal{C} \times (\mathcal{T}_C \times (\mathcal{I} \cup \{*\}))) \cup (\mathcal{R} \times \mathcal{T}_\mathcal{R})$ as follows:

$$\text{compatible}(t, t') = \text{true} :\Leftrightarrow \text{type}(t) = \text{type}(t') \tag{24}$$

In order to cushion this notational load, we introduce a simplified representation for SG. Given a dependency $\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}}$ involving the head (resp. the dependent) concept $\Theta^{a,b}_{i,j}$ (resp. $\Theta^{c,d}_{i,k}$), a relation $[b, \tau, d]$, and the corresponding edge $([b, \tau, d], 1, \Theta^{a,b}_{i,j})$ (resp. $([b, \tau, d], 2, \Theta^{c,d}_{i,k})$), it is summarized in the graph shown in Fig. 5. As example, Fig. 6 shows the SG for the sentence whose semantic is described in Fig 4. To facilitate better global understanding we do not make the indexes corresponding to either the number of the sentence in the text or the position of the form in that sentence explicit.

Figure 6: The SG of a sentence

6. The testing frame

In order to estimate the impact of our KE proposal as support for KD, our choice of testing frame fell on the IR one. We avoid some strategies in semantic indexing because we want to isolate the effects of our work in relation to the BOW models we use as baseline. So, we leave aside the consideration of external ontologies, often used to incorporate semantic relations through query expansion or term/relation compatibility, as well as the analysis of co-reference by anaphora. Our aim is to follow the guidelines outlined by TREC benchmarking exercises (`http://trec.nist.gov/`), which have standardized the use of *query relevance judgments* (QREL), as the heart of such a challenge. We then need a document collection, a set of topics and a set of trustworthy evaluation measures.

6.1. Selecting the evaluation metrics

We consider two groups of metrics: set and rank-based ones. In the former case, evaluation focuses on the relevant or non-relevant character of the documents retrieved, including *precision* and *recall*, as well as F and *fall-out* measures. These latter allow us to estimate the harmonic mean of precision and recall and to take into account the proportion of non-relevant documents retrieved, respectively. The second purpose takes into account the order in which the returned documents are presented. We here consider a wide range of metrics, starting with *precision* (resp. *recall*) *at k documents* (P@k) (resp. R@k), which permits us to compute these parameters even when we are only interested in fixed low levels of retrieved results as it is typically the case (Granka

et al., 2004). To this respect, we consider a restriction on the top 10 documents (P@10, R@10), determined by the approximate size of first page of results returned. The geometric interpretation of the precision-recall graph corresponds to the *mean average precision* (MAP). In order to highlight improvements for low-performing queries, we calculate the *geometric mean average precision* (GMAP). The *binary preference relation* (BPREF) distinguishes between documents that are explicitly judged as non-relevant and those that are only assumed to be non-relevant because they are unjudged. The *normalized discounted cumulative gain* (NDCG) evaluates separately the performance at each relevance level, penalizing the appearance of highly relevant documents lower down in a search result. All these measures are calculated in exact accordance with the TREC protocol with one proviso: the group of human experts does not come from the *National Institute of Standards and Technology* but from the Spanish *Centro Superior de Investigaciones Científicas* (CSIC).

6.2. Selecting the document collection

Although TREC proposes a number of domain-specific tracks (chemical, genomics, medical records and legal ones), to the best of our knowledge never before has a proposal on biodiversity data been considered. The same can be said of the rest of freely available testing collections such as CLEF (`http://www.clef-initiative.eu`), the initiative on information access systems with an emphasis on multilingual and multimodal information, or BIONLP (`http://www.bionlp.org`), the event on biology resources that focus on genomics and medical knowledge. Thus, we decided to choose the *corpus* $\mathscr{B}$, a real world compendium of botany, as our testing resource.

6.3. Selecting the topic set

Since we are far from classical tracks, we also need to define a selecting strategy for the topic set. Following the parameters of TREC events (Webber et al., 2008), we select 150 topics, although 50 are usually sufficient. The complete list of these topics can be consulted in the appendix.

6.3.1. The sampling process

We take the difficulty in solving queries, a major factor when both seeking to discriminate between IR systems (Mizzaro and Robertson, 2007) and justify the consideration of linguistically motivated retrieval (Egozi et al., 2008), as a dependent variable for sampling. So, we limit our choice to queries referring to narrative contents (Sparck Jones, 2000) and we classify the sample space (population) following two independent criteria correlated with difficulty, each creating its own partition:

- The *length* of the query is accepted as a factor of refinement (Phan et al., 2007) and long queries frequently contain extraneous terms that hinder the retrieval of relevant documents (Kumaran and Carvalho, 2009). New constraints have formalized the interaction between query-length and document-length normalization (Cummins and O'Riordan, 2012) as a factor impacting BOW retrieval.

- The *density* of the head-dependent graph associated to the query, i.e., the ratio between the number of edges and the number of possible edges (Coleman and More, 1983). It is proportional to both the computational effort needed to generate the knowledge representation and the specificity of the latter. In the former (resp. second) case, quantified by the number of input (resp. output) edges in a vertex, which determines the degree of ambiguity to be solved (resp. the number of modifiers for a given head). Density is defined in the interval [0-1].

In order to balance the sample, we minimize (resp. maximize) variability within (resp. between) sub-populations (strata) corresponding to different partitions. So, we uniformly distribute it between the sub-populations introduced for each partition, which provides homogeneity in all the levels of that stratification. Also, topics in a given stratum of a partition are equitably shared between the strata of the other one. We thereby ensure that the probability of one of the topics in the sample having a given length and density is approximately the same, regardless of the combination considered.

		Density				Length	
		[0.038-0.064]	[0.065-0.077]	[0.078-0.143]	**[0.038-0.143]**	Avg. val.	Std. dev.
		Number of queries					
Length	[2-7]	12	17	14	**43**	5.186	1.313
	[8-10]	19	19	19	**57**	9.052	0.788
	[11-17]	19	16	15	**50**	12.3	1.446
	[2-17]	**50**	**52**	**48**	**150**	**9.026**	**3.043**
Density	Avg. val.	0.060	0.071	0.095	**0.075**		
	Std. dev.	0.022	0.003	0.0177	**0.022**		

Table 6: Sample query distribution

In the case of length, we consider three sub-populations: short [2-7], medium [8-10] and long [11-17]. The number of queries in these intervals and the percentage of the total sample are, respectively: 43 (∼28.666%), 57 (∼38%) and 50 (∼33.333%). The mean average (resp. standard deviation) values are, also respectively: 5.186 (resp. 1.313), 9.052 (resp. 0.788) and 12.3 (resp. 1.446). The total mean average length is 9.026, with a standard deviation of 3.043, which means a range of lengths commonly considered as long (Bendersky and Croft, 2009) (short [1-4] and long [5-12]).

Where density is concerned, we also consider three sub-populations: low [0.038-0.064], medium [0.065-0.077] and high [0.078-0.143]. The number of queries in these intervals and the percentage of the total sample are, respectively: 50 (∼33.333%), 52 (∼34.666%) and 48 (32%). The mean average (resp. standard deviation) values are, also respectively: 0.060 (resp. 0.022), 0.071 (resp. 0.003) and 0.095 (resp. 0.017). The total mean average density is 0.075, with a standard deviation of 0.022. An overview of this sample query distribution on both length and density is compiled in Table 6.

6.3.2. Judging difficulty

We here follow (Mizzaro and Robertson, 2007), which shows that a system that aims to obtain good results in TREC needs to be effective on easy topics, although its real retrieval capability should be perphaps related to the treatment of difficult ones.

Definition 16. *Let $\sigma = \{\sigma_i\}_{i\in I}$ be a set of* IR *systems, $\mathcal{D}$ be a document collection and $\mathcal{Q} = \{\mathcal{Q}_j\}_{j\in J}$ be a topic (query). We then define the* average average precision of the set of IR systems σ on the topic $\mathcal{Q}_j$ for $\mathcal{D}$, *as:*

$$\text{AAP}(\sigma, \mathcal{Q}_j, \mathcal{D}) := \frac{\sum_{i\in I} \text{AP}(\sigma_i, \mathcal{Q}_j, \mathcal{D})}{|\sigma|}, \quad \textit{with } \text{AP}(\sigma_i, \mathcal{Q}_j, \mathcal{D}) \textit{ the average precision} \tag{25}$$

Intuitively, AAP is an indicator of the ease in topic satisfaction, understood as a magnitude related with the number of systems having good performance on that topic.

Definition 17. *Let $\sigma = \{\sigma_i\}_{i\in I}$ be a set of* IR *systems, $\mathcal{D}$ be a document collection and $\mathcal{Q} = \{\mathcal{Q}_j\}_{j\in J}$ be a set of topics (queries). We then define the* normalized average precision of σ_i on the topic $\mathcal{Q}_j$ according to $\text{AAP}(\sigma, \mathcal{Q}_j, \mathcal{D})$, *as:*

$$\text{NAP}_{\text{AAP}}(\sigma_i, \mathcal{Q}_j, \mathcal{D}) := \text{AP}(\sigma_i, \mathcal{Q}_j, \mathcal{D}) - \text{AAP}(\sigma, \mathcal{Q}_j, \mathcal{D}) \tag{26}$$

The adjacency matrix $[\text{NAP}_{\text{AAP}}(\sigma_i, \mathcal{Q}_j, \mathcal{D})]_{(i,j)\in I\times J}$ can be interpreted as a weighted bipartite graph, where the weight on arcs $\mathcal{Q}_j \rightarrow \sigma_i$ corresponds to the values for $\text{NAP}_{\text{AAP}}(\sigma_i, \mathcal{Q}_j, \mathcal{D})$, reflecting the performance of σ_i on the topic $\mathcal{Q}_j$ and eliminating the deviations due to topic ease (Wu and McClean, 2006). We can analyze it on the basis of the *Kleinberg's hits algorithm* (Kleinberg, 1999), using the *hubness* and *authority* indicators for locating high-quality information related to link structures.

Definition 18. *Let $\sigma = \{\sigma_i\}_{i\in I}$ be a set of* IR *systems, $\mathcal{D}$ be a document collection and $\mathcal{Q} = \{\mathcal{Q}_j\}_{j\in J}$ be a set of topics (queries). We define the* authority of the IR system σ_i on topics $\mathcal{Q}$ *(resp. the* hubness of the topic $\mathcal{Q}_j$ on IR systems σ*)* for $\mathcal{D}$, *as:*

$$\text{A}(\sigma_i, \mathcal{Q}, \mathcal{D}) := \sum_{j\in J} \text{H}(\mathcal{Q}_j, \sigma, \mathcal{D}).\text{NAP}_{\text{AAP}}(\sigma_i, \mathcal{Q}_j, \mathcal{D}) \tag{27}$$

$$(\textit{resp. } \text{H}(\mathcal{Q}_j, \sigma, \mathcal{D}) := \sum_{i\in I} \text{A}(\sigma_i, \mathcal{Q}, \mathcal{D}).\text{NAP}_{\text{AAP}}(\sigma_i, \mathcal{Q}_j, \mathcal{D})) \tag{28}$$

An IR system has high authority if it is more effective on topics with high hubness, in other words on difficult topics. Thus, we extract three subsets of 50 topics each from the initial sample on the basis of the degree of difficulty observed: high, medium and low.

6.4. Parameter tuning and baseline selection

We chose a sample of some of the most well-known and efficient BOW-based ranking functions to serve as reference for our proposal, which we have baptized as COGIR. Most of these functions have parameters that need to be tuned in order to obtain the

best results. Because our intention is to explore the performance at different levels of difficulty and the latter are estimated from those results as just noted, we are talking about two interdependent processes. To provide a reasonable exit from this problem we first tune each of the BOW-based ranking functions on all the topic set, which allows us to assign each query to its corresponding level of difficulty regardless of COGIR, the proposal we want to evaluate. We then optimize the tuning for each level and retrieval function, now including COGIR. The list of BOW-based ranking functions considered and the corresponding tuned parameters on the complete topic sample, taking MAP as reference and denoted by c super-indexes, is the following:

1. As algebraic distances, the pivoted cosine (Singhal et al., 1996) and the impact-based ranking (Anh and Moffat, 2002), both using a TF-IDF weighting factor. We tuned the *slope* in the pivoted cosine from 0 to 0.44 in increments of 0.04, to take the value $slope^c = 0.44$, while the authors suggest 0.2.
2. As probabilistic ranking, the Okapi's BM25 (Jones et al., 2000). We tuned b from 0 to 1 in increments of 0.05, obtaining $b^c = 0.3$. Since it seems that k_1 and k_3 have little effect on retrieval performance, we fix them to 1.2 and 1,000. Usually (He and Ounis, 2005) they are fixed to $k_1 = 1.2$, $k_3 = 1,000$ and $b = 0.75$.
3. As language model measure, the Dirichlet Smoothing (Zhai and Lafferty, 2004), for which we tune the μ parameter from 1,000 to 3,000 in increments of 100, resulting in the value $\mu^c = 2,800$. Following the authors, although the optimal prior depends on the collection, in most cases it is around 2,000.

The next step is to compute the authority (resp. hubness) for each BOW-based ranking function (resp. topic), which allows us to classify the topics according to their level of difficulty (hubness): low, medium and high levels lie within intervals $(0, 0.044]$, $(0.044, 0.076]$ and $(0.076, \infty)$, respectively. We now adjust the tuning on each interval, which includes 50 topics, using the same sweeping techniques as above:

1. The *slope* for the pivoted cosine takes the values $slope^h = 0.44$, $slope^m = 0.44$ and $slope^l = 0.44$ for the high, medium and low levels of difficulty, respectively.
2. The parameter b of Okapi's BM25 takes the values $b^h = 0.05$, $b^m = 0.3$ and $b^l = 0.2$ for the high, medium and low levels of difficulty, respectively. Parameters k_1 (resp. k_3) are fixed to values $k_1^h = 1.2$, $k_1^m = 1.2$ and $k_1^l = 1.2$ (resp. $k_3^h = 1,000$, $k_3^m = 1,000$ and $k_3^l = 1,000$) for high, medium and low levels, respectively.
3. The parameter μ of Dirichlet Smoothing takes the values $\mu^h = 2,800$, $\mu^m = 1,000$ and $\mu^l = 1,700$ for the high, medium and low levels, respectively.

We now take as a baseline among all the BOW-based ranking functions considered, for each metric and difficulty level, the one obtaining the best results, as shown in Table 8 by the underscored values. The implementation platform we have chosen to work with is ZETTAIR (`http://www.seg.rmit.edu.au/zettair/`). As both ZETTAIR and COGIR are written in C, this allows us to minimize the impact of implementation features on the tests.

As a general parameter setting for COGIR, we take $\upsilon_{to} = 0.7$, $\upsilon_{dto} = 0.7$ and $\upsilon_{dte} = 0.8$ as thresholds for the categorization processes described in Tables 1, 2 and 5, respectively. The maximum number of iterations is fixed in all cases (ι_{to}, ι_{dto} and ι_{dte}) to 10.

Taking MAP as a reference, we select the ratios μ_j and μ_a from 0 to 0.5 in increments of 0.05, obtaining $\mu_j^h = 0.2$ (resp. $\mu_a^h = 0.3$), $\mu_j^m = 0.2$ (resp. $\mu_a^m = 0.3$) and $\mu_j^l = 0.3$ (resp. $\mu_a^l = 0.2$) for high, medium and low levels of difficulty, respectively.

7. Experimental results

We can now input, visualize and interpret the results according to the different evaluation metrics and levels of difficulty considered for the topics, which we have condensed in Tables 7 and 8. We use bold (resp. underlined) fonts to mark the best overall values (resp. the baselines). Each value associates in brackets the percentage of improvement with regard to the corresponding baseline, reporting its statistical significance with respect to the latter ($p < 0.05$) using the Wilcoxon matched-pairs signed-ranks test (Wilcoxon, 1945) and marking it with a star. What is especially striking in both set and ranked-based results is the reduced range of the values obtained for the evaluation metrics, which reveals a non-trivial retrieval task.

7.1. Set-based evaluation results

These are summarized in Table 7, favoring COGIR over the other IR systems and reaching, with the exception of recall on the medium category (42.11%), the most significant percentage for all metrics on low difficult queries: precision (123.45%), F-measure (88.24%) and fall-out (-71.31%). With respect to the BOW models, all of them produce similar results on all grades of difficulty and metrics, with some differences that seem irrelevant against those previously mentioned for the conceptual approach[2]. The best (resp. the worst) absolute increase in the percentage for all ranking metrics is reached by COGIR for the precision (resp. the recall) on low (resp. on high) difficulty queries. We need to remember that the fall-out is a negative measure, in the sense that the best results are associated to minimum values.

	Precision	Recall	F-measure	Fall-out
LOW				
COGIR	**0.3620** (123.45%)*	**0.5674** (33.09%)*	**0.2210** (88.24%)*	**0.0788** (-71.31%)*
BM25	0.1620	0.4263	0.1174	0.2892 (5.27%)
DIRICHLET	0.1494 (-7.77%)	0.3932 (-7.76%)	0.1082 (-7.83%)	0.2891 (5.24%)
IMPACT	0.1602 (-1.11%)	0.4241 (-0.51%)	0.1163 (-0.93%)	0.2747
PIVOTED-COSINE	0.1616 (-0.24%)	0.4252 (-0.25%)	0.1171 (-0.25%)	0.2892 (5.27%)
MEDIUM				
COGIR	**0.3762** (115.95%)*	**0.6344** (42.11%)*	**0.2361** (87.97%)*	**0.0944** (-63.98%)*
BM25	0.1710 (-1.83%)	0.4341 (-2.75%)	0.1227(-2.30%)	0.2903 (10.75%)
DIRICHLET	0.1549 (-11.07%)	0.3933 (-11.89%)	0.1111 (-11.54%)	0.2904 (10.79%)
IMPACT	0.1748	0.4464	0.1256	0.2621
PIVOTED-COSINE	0.1682 (-3.44%)	0.4270 (-4.34%)	0.1207 (-3.90%)	0.2904 (10.79%)
HIGH				
COGIR	**0.3485** (104.04%)*	**0.6129** (29.52%)*	**0.2222** (77.19%)*	**0.1137** (-53.83%)*
BM25	0.1696 (-0.70%)	0.4690 (-0.88%)	0.1245 (-0.71%)	0.2929 (18.92%)
DIRICHLET	0.1508 (-11.71%)	0.4170 (-11.87%)	0.1107 (-11.72%)	0.2929 (18.92%)
IMPACT	0.1676 (-1.87%)	0.4635 (-2.05%)	0.1231 (-1.83%)	0.2463
PIVOTED-COSINE	0.1708	0.4723	0.1254	0.2929 (18.92%)

Table 7: Results on set-based evaluation measures

Except for fall-out, where the minimum score corresponds to the low difficulty case (0.7130), COGIR obtains the best numerical performance on medium difficult queries:

[2]the minimum increase for COGIR in relation to the baseline (29.52%) corresponds to recall with high difficulty, while the minimum decrease for BOW approaches (-0.24%) relies to the pivoted-cosine on precision.

precision (0.3762), recall (0.6344), F-measure (0.2631). The same reasoning can be applied for BM25: precision (0.1710), recall (0.4341), F-measure (0.1227), with minimum fall-out on easy queries (0.2892). The pivoted cosine behaved better with high difficulty sentences with the exception of fall-out, which once again reaches its minimum value at the easier category (0.2892). With respect to Dirichlet Smoothing (resp. to impact-based) ranking, we can observe a more complex behavior, with better values being obtained in the medium difficulty case for precision (0.1549) (resp. 0.1748) and F-measure (0.1111) (resp. 0.1256), while recall obtains them on high difficulty queries (0.4635) (resp. 0.4723). In relation to fall-out, Dirichlet Smoothing reaches the minimum value in the low difficulty case (0.2891), while impact-based ranking does so in that of the highest difficulty (0.2463). The best (resp. the worst) absolute numerical value with COGIR is attained by recall (resp. the F-measure) (0.6344) (resp. 0.2210) at the medium (resp. the low) level of difficulty.

7.2. Ranked-based evaluation results

These are compiled in Table 8, and corroborate the first impression obtained from the set-based approach, namely that the conceptual strategy better exploits the semantic relations which make up the meaning of the text[3]. However, the observed behavior for COGIR seems to be much more complex and the best percentages of improvement are now distributed between the three categories of difficulty: MAP (130.82%), BPREF (110.63%) and P@10 (62.87%) at the high level, GMAP (111.97%) at the medium level; and R@10 (67.00%) and NDCG (77.58%) at the low one. On the contrary, the worst performances correspond to the low level of difficulty for BPREF (95.74%) and P@10 (45.20%), the medium one for MAP (117.22%) and R@10 (58.09%), and the high level for GMAP (89.27%) and NDCG (63.11%). The best (resp. the worst) absolute increase in the percentage for all ranking functions is reached by COGIR for the MAP (130.82%) (resp. the P@10 (45.20%)) metric at the highest (resp. lowest) level of difficulty.

	MAP	GMAP	BPREF	P@10	R@10	NDCG
LOW						
COGIR	**0.4575** (129.43%)*	**0.3286** (100.36%)*	**0.4549** (95.74%)*	**0.5082** (45.20%)*	**0.3634** (67.00%)*	**0.7130** (77.58%)*
BM25	0.1967 (-1.35%)	0.1579 (-3.72%)	0.2169 (-6,67%)	0.3080 (-12.00%)	0.1872 (-13.97%)	0.4015
DIRICHLET	0.1942 (-2.60%)	0.1453 (-11.40%)	0.2288 (-1.54%)	0.3280 (-6.28%)	0.2088 (-4.04%)	0.3884 (-3.26%)
IMPACT	0.1717 (-13.89%)	0.0841 (-48.72%)	0.1675 (-27.92%)	0.2160 (-38.28%)	0.1463 (-32.76%)	0.3505 (-12.70%)
PIVOTED-COSINE	0.1994	0.1640	0.2324	0.3500	0.2176	0.4008 (-0.17%)
MEDIUM						
COGIR	**0.5574** (117.22%)*	**0.3860** (111.97%)*	**0.5555** (99.39%)*	**0.5388** (48.84%)*	**0.4116** (58.06%)*	**0.7812** (75.70%)*
BM25	0.2566	0.1821	0.2755 (-1.11%)	0.3460 -4.42%)	0.2533 (-2.72%)	0.4446
DIRICHLET	0.2380 (-7.24%)	0.1576 (-13.45%)	0.2618 (-6.03%)	0.3340 (-7.73%)	0.2533 (-2.72%)	0.4195 (-5.64%)
IMPACT	0.2166 (-15.58%)	0.0743 (-59.19%)	0.2117 (-24.01%)	0.2740 (-24.30%)	0.2187 (-16.01%)	0.3939 (-11.40%)
PIVOTED-COSINE	0.2546 (-0.77%)	0.1788 (-1.81%)	0.2786	0.3620	0.2604	0.4422 (-0.54%)
HIGH						
COGIR	**0.5219*** (-130.82%)	**0.3301*** (89.27%)	**0.5209*** (110.63%)	**0.5896*** (62.87%)	**0.4018*** (58.06%)	**0.7058*** (63.11%)
BM25	0.2109 (-6.72%)	0.1618 (-7.22%)	0.2343 (-5.25%)	0.3220 (-11.05%)	0.2444 (-3.85%)	0.4138 (-4.36%)
DIRICHLET	0.1960 (-13.31%)	0.1455 (-16.57%)	0.2139 (-13.54%)	0.3180 (-12.15%)	0.2393 (-5.86%)	0.3863 (-10.72%)
IMPACT	0.1735 (-23.26%)	0.0579 (-66.80%)	0.1759 (-28.87%)	0.2320 (-35.91%)	0.1528 (-39.89%)	0.3441 (-20.47%)
PIVOTED-COSINE	0.226	0.1744	0.2473	0.3620	0.2542	0.4327

Table 8: Results on rank-based evaluation measures

The conceptual model once again obtains the best numerical results on medium difficulty queries except for P@10, where the maximum value is reached in the high

[3]the minimum increase for COGIR in relation to the baseline (45.20%) corresponds to P@10 at low level of difficulty, while the minimum decrease for BOW approaches (-0.17%) relaties to the pivoted-cosine on NDCG.

case. The same applies to BM25, although here P@10 also shows a greater performance at the medium level. With respect to the rest of BOW models, all metrics reach their best value on medium difficulty queries. The worst numerical values correspond to the low level of difficulty for MAP (0.4575), GMAP (0.3286), BPREF (0.4549), P@10 (0.5082) and R@10 (0.3634); and the high one for NDCG (0.7058). The best (resp. the worst) absolute numerical value with COGIR is reached by the NDCG (resp. the GMAP) (0.7812) (resp. 0.3286) at the medium (resp. the low) level of difficulty.

All of the ranking functions achieve their best percentages (resp. numerical values) for the MAP (resp. NDCG) metric at all levels of difficulty, which suggest that documents are successfully evaluated at each relevance level. Curiously, all the minimum percentages (resp. numerical values) also concur on the P@10 (resp. GMAP) measure with the exception of R@10 at the most complex level of difficulty, demonstrating the complexity of highlighting improvements for precision and recall for the first page of returned results (resp. for low-performing topics). The minimum percentages (resp. numerical values) concur on the P@10 (resp. GMAP) for low and medium categories (resp. for all categories) of difficulty. At the highest level P@10 is only reduced by R@10, once again revealing the complexity of highlighting improvements for precision and recall for the first page of returned results (resp. for low-performing topics).

7.3. Interpreting results

The data reported above allow us to argue that COGIR performs significantly better than BOW-based IR regardless of the level of query difficulty, but particularly when this is high, while the percentage improvement seems to be more moderate at a medium level and even a little more restrictive at a low one. This is consistent with the fact that more difficult queries should require a larger amount of semantic information in order to produce a successful solution. From our perspective, since COGIR relaties heavily on the KD process described, these results are a practical confirmation of its goodness.

8. Conclusions

Systematics, the study of biological diversity and its evolution, uses taxonomy as a primary tool in understanding organisms, a process that involves the gathering and filtering of vast amounts of data mainly from scientific literature, most of which is currently available in digital format or being digitized. This can be modelled as a KD task in which TM is necessary to perform KE from raw texts.

We describe a proposal for TM that introduces KE from the semantic captured through the output of a dependency parse. The use of SG as representation structure for the meaning provides a platform upon which reasoning is acomplished on the perspective of graph matching, whilst ensuring the soundness and completeness of the process in FOL. This makes it possible to calculate the proximity between contents, locating the query answering in a decidable framework, a feature that enables the proposal to be exhaustively evaluated on the basis of its IR capabilities. The results suggest a significant uplift in performance, particularly when complex semantic relations are at stake.

Acknowledgments

Partially funded by the Ministry of Science and Innovation, and FEDER through project TIN2010-18552-C03-01, and by the Autonomous Government of Galicia through project CN 2012/317.

References

Anh, V., Moffat, A., 2002. Impact transformation: effective and efficient web retrieval. In: Proc. of the 25th Int. Conf. on Research and Development in Information Retrieval. SIGIR'02. Tampere, Finland, pp. 3–10.

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P. (Eds.), 2003. The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press.

Baget, J., Mugnier, M., Jun. 2002. Extensions of simple conceptual graphs: The complexity of rules and constraints. Journal of Artificial Intelligence Research 16 (1), 425–465.

Bendersky, M., Croft, W., 2009. Analysis of long queries in a large scale search log. In: Proc. of the 2009 Workshop on Web Search Click Data. WSCD'09. pp. 8–14.

Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., Salakoski, T., 2009. Extracting complex biological events with rich graph-based feature sets. In: Proc. of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. BioNLP'09. pp. 10–18.

Brown, J., Burton, R., 1975. Multiple representations of knowledge for tutorial reasoning. In: Bobrow, D., Collins, A. (Eds.), Representation and Understanding. Academic Press, New York, pp. 311–350.

Candito, M., Kahane, S., 1998. Can the TAG derivation tree represent a semantic graph? In: Proc. of 4th Int. Workshop on Tree Adjoining Grammars and Related Formalisms. TAG+4. pp. 21–24.

Clarke, C., Cormack, G., Burkowski, F., 1995. An algebra for structured text search and a framework for its implementation. The Computer Journal 38 (1), 43–56.

Coleman, T., More, J., 1983. Estimation of sparse jacobian matrices and graph coloring problems. SIAM Journal of Numerical Analysis 20 (1).

Covington, M., 1990. Parsing discontinuous constituents in dependency grammar. Computational Linguistics 16 (4), 234–236.

Creary, L., Pollard, C., 1985. A computational semantics for natural language. In: Proc. of the 23rd Annual Meeting on Association for Computational Linguistics. ACL'85. pp. 172–179.

Cummins, R., O'Riordan, C., 2012. A constraint to automatically regulate document-length normalisation. In: Proc. of the 21st Int. Conf. on Information and Knowledge Management. CIKM'12. pp. 2443–2446.

Daltio, J., Medeiros, C. B., Nov. 2008. Aondê: An ontology web service for interoperability across biodiversity applications. Information Systems 33 (7-8), 724–753.

de la Clergerie, E., 2010. Building factorized TAGs with meta-grammars. In: Proc. of 10th Int. Workshop on Tree Adjoining Grammars and Related Formalisms. TAG+10. pp. 111–118.

Deans, A., Yoder, M., Balhoff, J., Feb. 2012. Time to change how we describe biodiversity. Trends in Ecology & Evolution 27 (2), 78–84.

Delteil, A., Faron, C., 2002. A graph-based knowledge representation language for concept description. In: Proc. of the 15th European Conf. on Artificial Intelligence. ECAI'02. pp. 297–301.

Egozi, O., Gabrilovich, E., Markovitch, S., 2008. Concept-based feature generation and selection for information retrieval. In: Proc. of the 23rd Int. Conf. on Artificial intelligence. AAAI'08. pp. 1132–1137.

Ereshefsky, M., 2007. The Poverty of the Linnaean Hierarchy. Cambridge University Press.

Friedman, C., Kra, P., Yu, H., Krauthammer, M., Rzhetsky, A., 2001. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Computer Applications in the Biosciences 17 (suppl_1), S74–82.

Fundel, K., Kuffner, R., Zimmer, R., 2007. RelEx: Relation extraction using dependency parse trees. Journal on Bioinformatics 23 (3), 365–371.

Gardent, C., Kallmeyer, L., 2003. Semantic construction in feature-based TAG. In: Proc. of the 10th Conf. of the European Chapter of the Association for Computational Linguistics. EACL'03. pp. 123–130.

Genest, D., Chein, M., 2005. A content-search information retrieval process based on conceptual graphs. Knowledge and Information Systems 8 (3), 292–309.

Goldstein, P., DeSalle, R., 2011. Integrating dna barcode data and taxonomic practice: Determination, discovery, and description. BioEssays 33 (2), 135–147.

Gómez, C., Weir, D., Carroll, J., 2009. Parsing mildly non-projective dependency structures. In: Proc. of the 12th Conf. of the European Chapter of the Association for Computational Linguistics. EACL'09. pp. 291–299.

Granka, L., Joachims, T., Gay, G., 2004. Eye-tracking analysis of user behavior in WWW search. In: Proc. of the 27th Int. Conf. on Research and Development in Information Retrieval. SIGIR'04. pp. 478–479.

Harris, Z., 1954. Distributional structure. Word 10 (2-3), 146–162.

He, B., Ounis, I., 2005. Term frequency normalisation tuning for bm25 and dfr models. In: Proc. of the 27th European Conf. on Advances in Information Retrieval Research. ECIR'05. pp. 200–214.

Horrocks, I., 2005. Applications of description logics: State of the art and research challenges. In: Proc. of the Fifth Int. Conf. on Computational Science. ICCS'05. pp. 78–90.

Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E., McCouch, S., Pujar, A., Reiser, L., Rhee, S., Sachs, M., Schaeffer, M., Stein, L., Stevens, P., Vincent, L., Ware, D., Zapata, F., 2005. Plant ontology (PO): a controlled vocabulary of plant structures and growth stages: Research articles. Comparative and Functional Genomics 6 (7-8), 388–397.

Jensen, L., Saric, J., Bork, P., 2006. Literature mining for the biologist: from information retrieval to biological discovery. Natural Reviews Genetics 7 (2), 119–129.

Jones, K. S., Walker, S., Robertson, S. E., 2000. A probabilistic model of information retrieval: development and comparative experiments. Information Processing & Management 36 (6), 779–808.

Joshi, A., 1969. Properties of formal grammars with mixed types of rules and their linguistic relevance. In: Proc. of the Third Int. Conf. on Computational linguistics. COLING'69. pp. 1–18.

Kleinberg, J., 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM 46, 604–632.

Kumaran, G., Carvalho, V., 2009. Reducing long queries using query quality predictors. In: Proc. of the 32nd Int. Conf. on Research and Development in Information Retrieval. SIGIR'09. pp. 564–571.

Levine, R., 2006. Phrase Structure Grammar, Head-Driven. John Wiley & Sons, Ltd.

Maron, M., Kuhns, J., 1960. On relevance, probabilistic indexing and information retrieval. Journal of the ACM 7 (3), 216–244.

Minsky, M., 1974. A framework for representing knowledge. Tech. rep., Cambridge, MA, USA.

Miotto, O., Tan, T. W., Brusic, V., 2005. Supporting the curation of biological databases with reusable text mining. Genome informatics. Int. Conf. on Genome Informatics 16 (2), 32–44.

Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T., Tsujii, J., 2006. Semantic retrieval for the accurate identification of relational concepts in massive textbases. In: Proc. of the 21st Int. Conf. on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. ACL'06. pp. 1017–1024.

Mizzaro, S., Robertson, S., 2007. Hits hits TREC: exploring IR evaluation results with network analysis. In: Proc. of the 30th Int. Conf. on Research and Development in Information Retrieval. SIGIR '07. pp. 479–486.

Mungall, C., Aug. 2004. Obol: integrating language and meaning in bio-ontologies. Comparative and Functional Genomics 5 (6-7), 509–520.

Nesson, R., Satta, G., Shieber, S., 2010. Complexity, parsing, and factorization of tree-local multi-component tree-adjoining grammar. Computational Linguistics 36 (3), 443–480.

Parapatics, P., Dittenbach, M., 2009. Patent claim decomposition for improved information extraction. In: Proc. of the 2nd Int. Workshop on Patent Information Retrieval. PaIR'09. pp. 33–36.

Pereira, F., Warren, D., 1980. Definite clause grammars for language analysis. A survey of the formalism and a comparison with augmented transition networks. Artificial Intelligence 13, 231–278.

Phan, N., Bailey, P., Wilkinson, R., 2007. Understanding the relationship of information need specificity to search query length. In: Proc. of the 30th Int. Conf. on Research and Development in Information Retrieval. SIGIR'07. pp. 709–710.

Richens, R., 1956. Preprogramming for mechanical translation. Machine Translation 3, 20–25.

Sagot, B., 2010. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In: Proc. of the 7th Conf. on Int. Language Resources and Evaluation. LREC'10.

Salton, G., Buckley, C., Aug. 1988. Term-weighting approaches in automatic text retrieval. Information Processing & Management 24 (5), 513–523.

Salton, G., Wong, A., Yang, C., 1975. A vector space model for automatic indexing. Communications of the ACM 18 (11), 613–620.

Schabes, Y., Abeille, A., Joshi, A., 1988. Parsing strategies with 'lexicalized' grammars: application to tree adjoining grammars. In: Proc. of the 12th Conf. on Computational Linguistics. COLING'88. pp. 578–583.

Sekimizu, T., Park, H., Tsujii, J., 1998. Identifying the interaction between genes and gene products based on frequently seen verbs in MEDLINE abstracts. Genome Informatics, 62–71.

Shah, P., Jensen, L., Boué, S., Bork, P., 2005. Extraction of transcript diversity from scientific literature. PLoS Computational Biology 1 (1).

Singhal, A., Buckley, C., Mitra, M., 1996. Pivoted document length normalization. In: Proc. of the 19th Int. Conf. on Research and Development in Information Retrieval. SIGIR'96. pp. 21–29.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L., Eilbeck, K., Ireland, A., Mungall, C., Leontis, N., Rocca, P., Ruttenberg, A., Sansone, S., Scheuermann, R., Shah, N., Whetzel, P., Lewis, S., 2007. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25 (11), 1251–1255.

Sowa, J., 1976. Conceptual graphs for a data base interface. IBM Journal of Research and Development 20, 336–357.

Sowa, J., 1984. Conceptual structures: information processing in mind and machine. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Sparck Jones, K., 2000. Further reflections on trec. Information Processing & Management 36 (1), 37–85.

Taylor, A., 95. Extracting knowledge from biological descriptions. In: Proc. of the Second Int. Conf. on Building and Sharing Very Large-Scale Knowledge Bases. KB&KS'95. pp. 114–119.

Tesnière, L., 1959. Elements de syntaxe structurale. Editions Klincksieck.

Webber, W., Moffat, A., Zobel, J., 2008. Statistical power in retrieval experimentation. In: Proc. of the 17th Int. Conf. on Information and Knowledge Management. CIKM'08. pp. 571–580.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bulletin 1 (6), 80–83.

Wilson, E., 2003. The encyclopedia of life. Trends in Ecology & Evolution 18 (2), 77–80.

Wu, S., McClean, S., 2006. Evaluation of system measures for incomplete relevance judgment in IR. In: Proc. of the 7th Int. Conf. on Flexible Query Answering Systems. pp. 245–256.

Zhai, C., Lafferty, J., 2004. A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems 22 (2), 179–214.

Zweigenbaum, P., Demner-Fushman, D., Yu, H., Cohen, K., 2007. Frontiers of biomedical text mining: current progress. Briefings in Bioinformatics 8 (5), 358–375.

A. The topic set

We here include the set of topics used in the evaluation task, organized in strata characterized by the level of difficulty.

A.1. Topics with low level of difficulty

1. Quelque chose de pubescent (”`Something pubescent`”)
2. Je cherche une plante avec un rachis d'une certaine texture (”`I am looking for a plant with a rachis of a certain texture`”)
3. Quelles sont les plantes avec un limbe de couleur? (”`Which are the plants with a colored limb?`”)
4. Les plantes avec un limbe de couleur et fleur d'une certaine texture (”`Plants with a colored limb and flower of a certain texture`”)
5. Je cherche quelque chose de relativement court (”`I am looking for something relatively short`”)
6. Je cherche des graines avec des arilles d'une certaine forme (”`I am looking for seeds with arils of a certain shape`”)
7. Quelles sont les plantes qui ont une partie courte? (”`Which are the plants that have a short part?`”)
8. Je veux savoir celles qui ont une partie longue? (”`I want to know those that have a long part?`”)
9. Elles doivent avoir quelque chose d'obtus (”`They must have something obtuse`”)
10. Quelles sont celles qui ont un organe charnu? (”`Which ones are those that have a fleshy organ?`”)
11. La plante qui a des pétales linéaires et quelque chose frêle (”`The plant that has linear petals and something frail`”)
12. Je cherche un organe cylindrique (”`I am looking for a cylindrical organ`”)
13. Je cherche un fruit ovoïde (”`I am looking for an ovoid fruit`”)

14. Quelles sont les parties qui sont grêles ou acuminées? (”Which are the parts that are slender or acuminate?”)
15. Elles doivent avoir quelque chose d’une certaine forme (”They must have something in a certain shape”)
16. Je cherche une plante qui a le pistil d’une certaine taille (”I am looking for a plant that has the pistil of a certain size”)
17. Quelles sont les plantes qui ont une partie d’une certaine taille? (”Which are the plants that have a part of a certain size?”)
18. Je cherche un fruit obtus (”I am looking for an obtuse fruit”)
19. Quelles sont celles qui ont un organe charnu ou un fruit obtus? (”Which ones are those that have a fleshy organ or an obtuse fruit?”)
20. Je cherche celles qui ont un fruit avec les lobes ciliés (”I am looking for those with a fruit with ciliate lobes”)
21. Corolle avec les organes ciliés (”Corolla with ciliate organs”)
22. Quelles sont les parties qui ont des rhizomes? (”Which are the parts that have rhizomes?”)
23. Je cherche ceux qui ont une fronde d’une certaine couleur (”I am looking for those that have a frond of a certain color”)
24. Je cherche des fougères avec des rhizomes d’une certaine texture (”I am looking for ferns with rhizomes of a certain texture”)
25. Je cherche une partie de la penne à une certaine position (”I am looking for a part of the pinna at a certain position”)
26. Je veux savoir celles qui ont des sépales latéraux d’une certaine couleur (”I want to know those that have lateral sepals of a certain color”)
27. Je cherche une inflorescence vivace avec une certaine texture (”I am looking for a perennial inflorescence with a certain texture”)
28. Je veux savoir quelles sont les fougères d’une certaine taille qui ont des lobes (”I want to know which are the ferns of a certain size that have lobes”)
29. Elles doivent avoir des dents asymétriques ou de certaine forme (”They must have dentate leaves that are asymmetrical or of a certain shape”)
30. Je cherche quelque chose d’étalée avec des lobes linéaires (”I am looking for something spread out with linear lobes”)
31. Fruit d’une certaine forme (”Fruit of a certain shape”)
32. Quelles sont les plantes qui ont certaines parties avec un limbe pubescent? (”Which are the plants that have certain parts with a pubescent limb?”)
33. Fougères terrestres avec quelque chose portant des écailles (”Terrestrial ferns with something having scales”)
34. Je cherche des parties basales ou basilaires (”I am looking for basal or basilar parts”)
35. Je cherche des couleurs blanchâtres (”I am looking for whitish colours”)
36. Sépales ou quelque chose d’autre jaune (”Sepals or something else yellow”)
37. La plantes qui a des anthères avec quelque chose long (”The plant that has anthers with something long”)

38. Quelles sont celles qui ont quelque chose d'alterne avec une partie acuminée? ("Which ones have something alternate with an acuminate part?")
39. Sore à indusie d'une certaine couleur et taille ("Sore with indusium of a certain color and size")
40. Quelque chose sessile et sigmoïde ("Something sessile and sigmoid")
41. La plante a un organe samaroïde ou linéaire ("The plant has a samaroid or linear organ")
42. Quelles sont celles qui ont un éperon d'une certaine forme ou spiciforme? ("Which ones have a spur of a certain shape or that is spicate?")
43. Les plantes qui ont les restes du rostelle de certaines formes ("Plants with the remnants of the rostellum of certain shapes")
44. Contrefort de certaine taille et forme ("Buttress of a certain size and shape")
45. Je cherche des organes médians ou très larges ("I am looking for medium-sized or very large organs")
46. Cette plante a des parties dentées ou acuminées ("This plant has dentate or acuminate parts")
47. Le limbe a quelques choses d'acuminés ("The limb has acuminate things")
48. Je veux savoir quelles sont celles qui ont une nerville portant une veinule à une position ("I want to know which ones have a veinlet having a venule at a position")
49. Je cherche quelque chose portant des écailles de certaines couleurs ("I am looking for something having scales of certain colors")

A.2. Topics with medium level of difficulty

1. Plantes avec stipules ("Plants with stipules")
2. Quelles sont les plantes qui ont des stipules persistantes? ("Which are the plants that have persistent stipules?")
3. Je cherche les plantes qui ont des bractées pubescentes ("I am looking for plants that have pubescent bracts")
4. Plantes avec des gousses longues de 14 cm ("Plants with 14-cm long pods")
5. Je cherche les plantes qui ont des graines noires ("I am looking for plants that have black seeds")
6. Quelles sont les plantes qui ont des pétales onguiculés? ("Which are the plants that have unguiculate petals?")
7. Plantes avec graine obovoïde ("Plants with obovoid seeds")
8. La plante a des feuilles obtuses ("The plant has obtuse leaves")
9. Limbe denté ou acuminé ("Dentate or acuminate limb")
10. Je veux savoir quelles sont celles qui ont des graines avec arilles ("I want to know which ones have seeds with arils")
11. Quelles sont les plantes qui ont des pinnules sur le costae canaliculées? ("Which are the plants that have pinnules on their canaliculate costae?")
12. Je veux savoir quelles sont les plantes qui ont un rhizome portant des écailles ("I want to know which plants have a rhizome carrying scales")

13. Quelles sont les plantes qui ont un pétiole long de 9 cm? (”Which are the plants that have a 9-cm long petiole?”)
14. Le sépale dorsal est mince (”The dorsal sepal is thin”)
15. Je veux savoir quelles plantes ont des sépales latéraux (”I want to know which plants have lateral sepals”)
16. Plantes avec des feuilles acuminées (”Plants with acuminate leaves”)
17. Plantes avec 1 inflorescence dense (”Plants with 1 dense inflorescence”)
18. Quelles sont les plantes qui ont des bractées florales? (”Which are the plants that have floral bracts?”)
19. Je cherche des feuilles avec des folioles elliptiques (”I am looking for leaves with elliptic leaflets”)
20. Quelles sont les plantes qui ont le limbe des feuilles coriace? (”Which are the plants that have a leathery-leaved limb?”)
21. Quelles sont celles avec des pétioles larges ou longs? (”Which ones are those with wide or long petioles?”)
22. Je cherche des plantes avec les pétales et feuilles falciformes (”I am looking for plants with falcate petals and leaves”)
23. Une gousse samaroïde ou linéaire (”A samaroid or linear pod”)
24. Plantes qui a un éperon cylindrique et spiciforme (”Plants that have a cylindrical and spicate spur”)
25. Le staminode ou la drupe est charnu (”The staminode or drupe is fleshy”)
26. Quelles sont les plantes qui ont un ovaire hirsute avec des ovules (”Which are the plants that have a hirsute ovary with ovules”)
27. Je cherche un rameau avec des ombelles circulaires (”I am looking for a branchlet with circular umbels”)
28. Quelles sont les plantes avec un calice et des glandes brillantes? (”Which are the plants with a calyx and shiny glands?”)
29. Je veux savoir quelles sont celles qui ont un calice avec des glandes et des périanthes cupuliformes (”I want to know which ones have a calyx with glands and cup-shaped perianths”)
30. La plante a un style falciforme ou glabre (”The plant has a falcate or glabrous style”)
31. Quelles sont celles qui ont des pennes latérales ou des pennes inférieures? (”Which ones are those that have lateral pinna or inferior pinna?”)
32. Le reste du rostelle est trilobé (”The remnant of the rostellum is trilobate”)
33. Ces plantes ont les tubes du calice verts (”These plants have green calyx tubes”)
34. La plante qui a des anthères avec des déhiscences longues (”The plant which has anthers with long dehiscences”)
35. Tubercule unique (”Single tubercle”)
36. Je cherche des gaines ou des nervures basales (”I am looking for basal sheaths or veins”)
37. Sépales ou tépales jaunes (”Yellow sepals or tepals”)

38. Je veux savoir quelles sont celles qui ont des étamines avec des anthères connectives (”I want to know which ones have stamens with connective anthers”)
39. Anthères avec valves transversales (”Anthers with transversal valves”)
40. Les plantes qui ont les aisselles des feuilles caduques (”Plants that have caducous leaf axils”)
41. Je veux savoir quelles ont les tubercules ellipsoïdes et uniques (”I want to know which ones have single ellipsoid tubercles”)
42. Cette plante a des contreforts ou les racines minces (”This plant has thin buttresses or roots”)
43. La plante a un tronc couvert d'écaille brune (”The plant has a trunk covered with brown scales”)
44. Un sore sur une nervure courte (”A sorus on a short vein”)
45. Je cherche un style à appendice uniflore (”I am looking for a style with a uniflorous appendix”)
46. Quelles sont les plante qui ont un limbe avec un lobe denté? (”Which are the plants that have a limb with a dentate lobe?”)
47. Le limbe a les lobes acuminés (”The limb has acuminate lobes”)
48. Une nerville portant une veinule circulaire (”A veinlet carrying a circular venule”)
49. Je cherche une plante qui a entre 12 - 14 ovules basales (”I am looking for a plant that has between 12 to 14 basal ovules”)
50. La plante a des racines portant des écailles foncées (”The plant has roots carrying dark scales”)

A.3. Topics with high level of difficulty

1. Rachis grêle (”Slender rachis”)
2. Plantes avec graine ovoïde (”Plant with ovoid seed”)
3. Quelles sont les plantes qui ont les tiges relativement courtes? (”Which ones are the plants that have relatively short stems?”)
4. Je veux savoir quelles sont les plantes qui ont les inflorescences relativement courtes (”I want to know which plants have relatively short inflorescences”)
5. Je cherche celles qui ont des gousses ligneuses très épaisses (”I'm looking for those that have very thick woody pods”)
6. Plantes avec un fût étroit et cylindrique (”Plants with a narrow cylindrical trunk”)
7. Quelles sont celles qui ont des feuilles oblongues ou oblongues-lancéolées? (”Which ones are those that have oblong or oblong-lanceolate leaves?”)
8. Les plantes qui ont des feuilles obtuses ou arrondies (”Plants that have obtuse or rounded leaves”)
9. Quelles sont les plantes qui ont un rachis grêle et pubescent? (”Which plants are those that have a slender and pubescent rachis?”)
10. Quelles sont les plantes qui ont des stipules velues et courtes? (”Which plants are those that have hairy and short stipules?”)

11. Quelles sont celles qui ont une graine avec des arilles jaunes? (”Which ones are those that have a seed with yellow arils?”)
12. Quelles sont les plantes qui ont des graines noires avec des arilles jaunes? (”Which plants have black seeds with yellow arils?”)
13. On cherche celles qui ont une corolle blanc ou rose (”We are looking for those that have a white or rose corolla”)
14. Quelle est celle qui a une graine obovoïde ou ovoïde? (”Which one has an obovoid or ovoid seed?”)
15. Quelles sont les plantes qui ont les étamines externes avec des anthères de 4 mm? (”Which plants have external stamen with 4-mm anthers?”)
16. Quelles sont les plantes qui ont des bractées florales membraneuses? (”Which plants have membranous floral bracts?”)
17. Quelles sont les plantes qui ont le labelle obtus ou ovale? (”Which plants have an obtuse or oval labellum?”)
18. Quelles sont les plantes qui ont un labelle avec des nervures épaisses? (”Which plants have a labellum with stout veins?”)
19. Quelles sont les plantes qui ont le pédicelle grêle et glabre? (”Which plants have a slender and glabrous pedicel?”)
20. La plante qui a des pétales minces et des sépales latéraux glabres (”The plant that has thin petals and glabrous lateral sepals”)
21. La plante qui a des pétales linéaires et des bractées courtes (”The plant that has linear petals and short bracts”)
22. Je cherche des feuilles alternes à nervures (”I am looking for alternate leaves with veins”)
23. Quelles sont les plantes qui ont un labelle avec des nervures pubescentes? (”Which plants have a labellum with pubescent veins?”)
24. Elles doivent avoir une gousse vive (”They must have a live pod”)
25. Je veux savoir quelles sont les plantes qui ont un rhizome portant une fleur en racème (”I want to know which plants have a rhizome with a flower in raceme”)
26. Je veux savoir quelles sont les plantes qui ont un arbrisseau portant des fleurs petites (”I want to know which plants have a treelet carrying little flowers”)
27. Quelles sont celles qui ont une corolle à lobes violets? (”Which ones have a corolla with violet lobes?”)
28. Fougères à rhizome petites (”Ferns with little rhizomes”)
29. Je cherche une plante avec limbe deltoïde et pétiole roussâtre (”I am looking for a plant with a deltoid limb and a reddish petiole”)
30. Plantes qui ont un rhizome portant des écailles obtuses avec des frondes (”Plants with a rhizome having obtuse scales with fronds”)
31. Je cherche celles avec un pétiole grisâtre et long de 9 cm (”Looking for those with a 9-cm long greyish petiole”)
32. Je veux savoir quelles sont celles qui ont des nervures espacées et bifurquées (”I want to know which ones have spaced and bifurcated veins”)

33. Je cherche des plantes avec des sépales latéraux linéaires (”I am looking for plants with linear lateral sepals”)
34. Plante qui a un pétiole straminé (”Plant which has a "stramineous" petiole”)
35. Plantes avec des feuilles acuminées avec les nervures épaisses (”Plant with acuminate leaves with thick veins”)
36. Elles doivent avoir des dents asymétriques (”They must have asymmetrical teeth”)
37. Pennes dorsales alternes (”Alternate dorsal pinna”)
38. Je cherche celles qui ont un ovaire hirsute et des ovules hispides (”I am looking for a hirsute ovary and hispid ovules”)
39. Je cherche des feuilles alternes avec des folioles elliptiques (”I’m looking for alternate leaves with elliptic leaflets”)
40. Tige étalée avec feuilles linéaires (”Spread stem with linear leaves”)
41. Quelles sont les plantes qui ont le limbe des feuilles sessiles coriace? (”Which plants a limb with coriaceous sessile leaves?”)
42. Fougères terrestres avec rhizome portant des écailles (”Terrestrial ferns with rhizome having scales”)
43. Quelles sont celles qui ont des sépales, des tépales ou des bractées jaunes? (”Which ones are those that have sepals, tepals or yellow bracts?”)
44. Elles doivent avoir les anthères ou les valves longues avec des déhiscences (”They must have anthers or long valves with dehiscences”)
45. Je veux celles qui ont le sore avec une indusie pâle et mince (”I want those that have a sorus with a light and thin indusium”)
46. Une fronde qui a des pennes mucronés portant des sporanges (”A frond which has mucronate pinna having sporangia”)
47. Cette plante a une indusie entière, membraneuse et pâle (”This plant has a whole, membranous and light-coloured indusium”)
48. Quelles sont celles qui ont un limbe à lobe denté ou acuminé? (”Which ones have a limb with dentate or acuminate lobes?”)
49. Ces plantes ont le foliole avec des lobes dentés ou acuminés (”These plants have leaflets with dentate or acuminate lobes”)
50. Ces plantes ont les fleurs roses avec des pseudonervures ligneuses (”These plants have rose flowers with woody pseudoveins”)