OpenLogos Semantico-Syntactic Knowledge-Rich Bilingual Dictionaries

Anabela Barreiro[1], Fernando Batista[1,2], Ricardo Ribeiro[1,2], Helena Moniz[1,3], Isabel Trancoso[1,4]

[1] L2F – INESC-ID, Rua Alves Redol 9, 1000-029, Lisboa, Portugal
[2] ISCTE-IUL - Instituto Universitário de Lisboa, Portugal
[3] FLUL/CLUL - Faculdade de Letras / Centro de Linguística - Universidade de Lisboa, Portugal
[4] IST - Instituto Superior Técnico - Universidade de Lisboa, Portugal
anabela.barreiro@inesc-id.pt

Abstract

This paper presents 3 sets of OpenLogos resources, namely the English-German, the English-French, and the English-Italian bilingual dictionaries. In addition to the usual information on part-of-speech, gender, and number for nouns, offered by most dictionaries currently available, OpenLogos bilingual dictionaries have some distinctive features that make them unique: they contain cross-language morphological information (inflectional and derivational), semantico-syntactic knowledge, indication of the head word in multiword units, information about whether a source word corresponds to an homograph, information about verb auxiliaries, alternate words (i.e., predicate or process nouns), causatives, reflexivity, verb aspect, among others. The focal point of the paper will be the semantico-syntactic knowledge that is important for disambiguation and translation precision. The resources are publicly available at the METANET platform for free use by the research community.

Keywords: Language Resources, Bilingual Dictionaries, OpenLogos, Semantico-Syntactic Knowledge

1. Introduction

Bilingual dictionaries represent a very important resource in machine translation. Even if translation does not consist in simply mapping words or multiword units between two languages, dictionaries represent the foundation of a translation system, and the more knowledge-richer they are, the more they can contribute to quality translation. The bilingual dictionaries that will be described in this paper have some features that make them unique. Unlike other dictionaries available, they contain semantico-syntactic and ontological relations, which have been developed inductively, by trial and error, over a period of years by the Logos MT system development team and used successfully for several decades in the Logos commercial machine translation product, now available as open source software under the name OpenLogos. OpenLogos uses a Semantico-Syntactic Abstraction Language called SAL, which represents both meaning (semantics), and structure (syntax). SAL is an interlingua-style hierarchical taxonomy comprising over 1,000 elements embracing all parts of speech, and distributed in supersets, sets and subsets, which are embedded in each dictionary entry and in the translation system's rules. This internal language was designed as an extensible system, so that developers would expand and add to its capabilities. It was initially developed for the English language, but many of its elements are universal and therefore applicable to other languages. Unlike other representation languages, SAL places semantics and syntax on a continuum, i.e., undissociated one from the other and represented in the same layer. SAL is eclectic in the categories included in the representation schema and it was designed to work in concert with other linguistic resources, namely lexical resources and a diverse set of linguistic rules, including the transfer (TRAN) and semantico-syntactic (SEMTAB) rules (Barreiro et al., 2011). We believe that these resources can be useful to enhance other machine translation systems, especially due to their semantico-syntactic knowledge richness. Furthermore, the abstraction echelon makes the ontology applicable at several levels and useful for natural language processing applications other than machine translation.

This paper is structured as follows: Section 2. describes the related work. Section 3. presents the most peculiar characteristics of the OpenLogos data. Section 4. highlights the semantico-syntactic knowledge embedded in the OpenLogos system. Section 5. shows the quantitative results. Finally, Section 6. presents the conclusions and future work.

2. Related Work

The effort on creating computerized bilingual lexicons goes back to the 1980s. (Picchi et al., 1988) describe the development of a bilingual lexical database system pairing English and Italian. At the same time, one of earliest European attempts on machine translation was running: the EUROTRA project (A. Raw and Eynde, 1988). Since then, there were several large research projects aiming at the development of specifications that would facilitate the reuse of linguistic resources or the development of such harmonized resources. A few examples are Genelex (Antoni-Lay et al., 1994), PAROLE/SIMPLE (Lenci et al., 2000), or Wordnet (Miller, 1995; Fellbaum, 1998). Framenet (Baker et al., 1998) has also similar objectives focusing in semantic knowledge. Green et al. (Green et al., 2004) further develop this idea by creating a system for automatically inducing frame semantic verb classes from Wordnet and the Longman Dictionary of Contemporary English (Procter, 1978). After this stage, the research focused on the development of resources that defined inter-language relations (Villegas et al., 2000). The MILE (Multilingual ISLE Lexical Entry) was one of the efforts based on previous resources to implement such relations (Calzolari et al., 2002). The

focus on the promotion of the reuse of existing resources gained great attention, with LREC 2004 (http://www.lrec-conf.org/lrec2004/) featuring a dedicated workshop to the subject: "A Registry of Linguistic Data Categories within an Integrated Language Resources Repository Area". Another interesting project focusing in creating inter-language links was EuroWordNet (Vossen, 1998).

The resources we present in this paper were used in the Logos commercial machine translation system during 2-3 decades. They have been validated by the former development team of this company and by its clients. We believe that they have unique characteristics and can be used as a basis for new linguistic and natural language processing tools, especially for poor-resourced languages.

3. OpenLogos Data

OpenLogos is the open source derivative of the Logos Machine Translation System developed by Bernard Scott in the early seventies (Scott, 2003; Barreiro et al., 2011). The strength of the Logos system resides in its lexical resources, the knowledge-rich bilingual dictionaries, which work in combination with distinct-purpose rulebases. OpenLogos dictionaries present several idiosyncrasies that distinguish them from other publicly available dictionaries. In addition to the most usual information on part-of-speech (POS), gender (GEN), and number (NUM) for nouns, OpenLogos bilingual dictionaries contain semantico-syntactic knowledge for all lexical entries. This information is represented at an abstract level by the SAL language (described in more detail in Section 4.), which contains 3 levels of representation: superset (SUPER), set (SET), and subset (SUB), and is used to help with disambiguation. OpenLogos lexicon also contains morphological information for source and target words. Because all words have morphological paradigms (PAT) assigned to them, it is possible to map inflected forms across languages (source and target). Morphological information is useful in improving agreement in machine translation. Another interesting feature of the dictionaries is the information about whether a source word is an homograph (HOMO) or not. Homographs are a major source of translation errors and their identification is crucial to help resolve those errors. Multiword units (henceforth, multiwords) contain information about the head word (HEAD). This knowledge can be used in generation and to correct machine translation problems related to agreement within multiword structures or within larger units, such as the agreement between nominal multiwords and the verb or agreement within verbal multiwords, such as in support verb constructions. Information about the auxiliary (AUX) of a verb is also provided to improve precision in the translation, especially in those cases where auxiliary choice is subtle. In addition, there is information about the alternate word (ALT), i.e., nominalization or process noun, predicate adjective, among others. This information can be used for paraphrasing purposes, for example, when the verb *present* is translated as the support verb construction *make a presentation* (*of*). Other information, such as whether a verb is causative (CAUS) or reflexive (REFL) in the target language, and aspectual information (ASP) for each verb is also available.

4. Semantico-Syntactic Knowledge

The most interesting aspect of the OpenLogos system is its ontology-based internal representation language: the Semantico-Syntactic Abstraction language (SAL). This representation language allows words to be represented at a higher level of semantic abstraction (a second order), and in many instances it allows disambiguation to take place at the lexical level. SAL permits easy mapping from natural language to symbolic language, representing both meaning (semantics), and structure (syntax) in a continuum. SAL was motivated by the belief that the semantics of a word often affects the surrounding syntax. For example, all verbs that call for indirect objects (di-transitive verbs) would appear to be semantically related (*send*, *communicate*, *convey*, *give*, *transmit*, *provide*, *supply*, etc.). Thus one can relate the syntactic effect of di-transitive verbs to their semantics. SAL was designed to subcategorize words according to these property/effect relationships. Thus, SAL seeks to capture the semantic properties of words having syntactic effect. For example, the nouns *resistance*, *relationship*, *marriage*, *accommodation*, all share a common SAL code reflecting their governance of the preposition *to* (*resistance to change*). The adjectives *easy*, *fun*, *simple* in (1) have SAL codes different from the codes for *eager*, *reluctant*, *determined* in (2), to reflect the pronounced differences in the syntactic effect of these groups. The loose semantic kinship is shared by members of each adjectival group.

(1) John is easy/fun/simple to please.

(1') Pleasing John is ?easy/fun/simple.

(2) John is eager/reluctant/determined to please.

(2') *Pleasing John is eager/reluctant/determined.

In the OpenLogos system, SAL knowledge is embedded in the dictionary in the form of numeric codes. For the sake of understandability, we use SAL mnemonics. For example, the noun (N) *table* has two SAL representations: one with the SAL code (COsurf), which contains the properties *concrete* and *surface*, and another one with the SAL code (INdata), which stands for *information*, *recorded data*. There are more than 1,000 SAL categories, organized in a hierarchical taxonomy of supersets, sets, and subsets, distributed by all parts-of-speech (POS). The complete list of SAL categories can be viewed at http://www.l2f.inesc-id.pt/∼abarreiro/openlogos-tutorial/new_A2menu.htm.

The existing elements of the SAL ontology are also documented in the SAL tutorial of the LearnLogos application that comes with the OpenLogos system, downloadable from the DFKI website or from SourceForge.

In the OpenLogos dictionaries, all POS categories are represented, contemplating variable and invariable words, multiwords and named entities. For example, the word *alligator* is classified as a noun (N) that inflects like the word *book* (PAT 16), where *book* is the word representing the morphological paradigm for regular nouns that take an *–s* to form the plural. Its SAL mnemonic (ANrept) stands for *animate*, *reptiles*. It designates cold-blooded, egg-laying vertebrates. The word *enter* is classified as a verb (V) that inflects like the verb *walk* (PAT 1), where

SAL id	Mnemonics	Description	Examples
8 46	MEabs	abstract measurable concepts	*humidity*, *lenght*
8 95	MEdis	discrete measurable concepts	*sum*, *increment*
8 61	MEunit	units of measure	See subsets
8 61 161	MEunitwt	units of weight	*ounce*, *pound*
8 61 162	MEunitvel	units of velocity	*mph*, *megahertz*
8 61 163	MEunitvol	unites of volume measure	*gallon*, *liter*
8 61 164	MEunittemp	units of temperature	*degrees celsius*
8 61 165	MEunitener	units of energy/force	*watt*, *horsepower*
8 61 234	MEunitsys	measurement systems	*fahrenheit*, *kelvin*
8 61 166	MEunitdur	units of duration	*hour*, *year*
8 61 167	MEunitspec	specialized units of measure	*oersted*, *ohm*
8 61 168	MEunitvalue	units of money/value	*dollar*, *euro*
8 61 170	MEunitlin	units of linear/area measure	*inch*, *mille*
8 61 169	MEundif	undifferentiated measure	*degree*, *share*

Table 1: Noun Measure

walk is the word representing the morphological paradigm for regular verbs like *walk*, *walked*, *walking*. Its SAL mnemonic (INMOIntoType) represents motional intransitive verbs (INMO), which comprise all verbs of motion, such as *depart*, *go*, *fly*, *run*, *walk*. This SAL group of verbs take kinetic-type prepositions such as *into*, *onto,* and *up to* denoting directed motion. The word *approximate* is classified as an adjective (A) that inflects like *natural* (PAT 34), where *natural* is the word representing the morphological paradigm for adjectives like *natural*, *more natural*, *most natural*. Its SAL mnemonic (AVquan) stands for adverbial adjectives of quantity/measure type. Quantity/measure type adjectives (i) denote notions of quantity or measure (*slight*, *extensive*); (ii) have an adverbial counterpart (*slightly*, *extensively*); (iii) may occur in the predicate adjective position (*the effect was extensive*); and (iv) do not govern prepositions. The word *yesterday* is classified as an adverb (ADV) that does not inflect. Its SAL mnemonic (TEMPpuncpast) stands for temporal adverbs that denote some aspect of time, answering the question *when*; punctual adverbs (punc) denote a point in time and (past) signifies past time (tense), such as *recently*, *previously*, or *a long time ago*. The word *several* is classified as an invariable pronoun (PRO), impersonal (IMPERS), indefinite (INDEF). The word *which* is classified as a relative and interrogative pronoun (RELINT). The word *or* is classified as a conjunction (CONJ), conjoining (JOIN). The word *alongside* is classified as a preposition (PREP), defined as locative (LOC), uninflected. The word *many* is classified as an invariable determiner (DET), plural (PL). Finally, the multiword *one third* is classified as an arithmate (ARITHM), numeric expression (NUM), fractional (FRAC). Sections 4.1., 4.2., and 4.3. describe SAL categories in detail for the 3 largest POS categories: verbs, nouns, and adjectives, respectively.

4.1. Nouns

Nouns (word class 01) have 12 supersets: concrete (CO), mass (MA), animate (AN), place (PL), information (IN), abstract (AB), process (intransitive) (PNin), process (transitive) (PNtr), measure (ME), time (TI), aspective (AS), and unknown (UN). All these supersets branch into their corresponding sets, and sometimes, subsets. Among the noun supersets, the superset measure (ME), for example, has 3 sets and 11 subsets, as illustrated in Table 1. All SAL codes for nouns represent semantic groupings, and is language independent, as concepts are transverse across languages.

4.2. Verbs

Verbs (word class 02) are subdivided in 3 main types: the intransitive, the weak transitive and the strong transitive. For example, the intransitive verbs have 3 distinct supersets: the motional (INMO), the existential (INEX), and the operational (INOP).

Motional intransitive verbs comprise all verbs of motion and include the INMOinto-type and the INMOin-type sets. The INMOinto-type includes verbs like *depart*, *go*, and *walk*. The INMOin-type includes verbs like *dance*, and *sail*. INMOinto-type verbs can take kinetic-type prepositions, such as *into*, *onto*, and *up to*, denoting directed motion. If the verbs of motion do not take these prepositions, they are classified as INMOin-type.

Existential intransitive verbs include the verb *be* and various *be*-substitute verbs that take predicate nominatives, such as *become* and *remain*, and predicate adjectives, such as *grow* and *sound*. Existential intransitives include the following 4 sets: INEXbe-type, INEXbecome-type, INEXgrow-type, and INEXseem-type. INEXbe-type includes verbs like *be*. INEXbecome-type includes verbs like *become* and *remain*. INEXgrow-type includes verbs like *grow*, *look*, and *sound*. INEXseem-type includes verbs like *appear* and *seem*. In general, pre-clausals are transitive, except for a small class of cases, such as *appear*, *seem*, and *insist*. Verbs like *agree* and *think* also have a non-pre-clausal intransitive function. OpenLogos parser selects between the transitive pre-clausal and the intransitive one.

Operational intransitive verbs denote all intransitive verbs that are not existential or verbs of motion. This includes intransitive verbs that take clausal and verbal complementation (except for *appear* and *seem*, which are existential intransitives). Operational intransitives include the following 5 sets: INOPmisc, INOPloc, INOPpcl, INOPprev, and INOPprecv. INOPmisc includes verbs like *sing*. INOPloc includes verbs like *stand*. INOPpcl includes verbs like *com-*

	SAL id	Mnemonic	Example Verb	Example Sentence
Existential	11 60	INEXbe-type	*be*	She *is* the valedictorian of her class. She *was* at the seashore all summer.
	11 61	INEXbecome-type	*become*, *remain*	He *became* a doctor at a very young age. He *remained* a Democrat all his life. She *remained* at the seashore all summer.
	11 64	INEXgrow-type	*sound*, *grow*, *look*	Their voices *sounded* cheerful. The day *grew* cooler.
	11 76	INEXseem-type	*seem*, *appear*	He *seemed* happy with the results. It *seems* that the operation was successful.
Operational	12 29/31/97	INOPmisc	*sing*	She *sings* well.
	12 68	INOPloc	*stand*	He *stood* in the rain.
	12 69	INOPpcl	*comment*	I *commented* about the parking problems.
	12 72	INOPprev	*refrain*, *persist*, *consist*, *conspire*, *participate*	They *refrained* from smoking. She *persisted* in pursuing her goals. The work *consisted* in checking papers. They *conspired* to defeat the candidate. He *participated* in solving the problem.
	12 73	INOPprecv	*insist*	He *insisted* on joining them. They *insisted* that the answer was correct.
Motional	10 24	INMOinto-type	*depart*, *walk*, *drive*, *go*	They *departed* for Chicago this morning. They *walked* into the room. They *drove* to the library.
	10 68	INMOin-type	*dance*, *sail*	They *danced* in the streets. They *sailed* around the lake all morning.

Table 2: Intransitive Verbs

ment. INOPprev includes verbs like *persist*, and INOPprecv includes verbs like *insist*. INOPloc verbs strongly claim locative prepositions (*He stayed at the office until midnight*).

Because of the endless richness of verb argument structures, the SAL verb taxonomy captures only salient features. The Logos model depends upon use of the Semantic Table (SEMTAB) to capture argument structures not provided for in the taxonomy. Table 2 illustrates the 3 different types of intransitive verbs. A full description of the transitive verbs (weak and strong) can be found in the SAL Tutorial.

4.3. Adjectives

Adjectives are classified in 2 types: descriptive and participial. For example, descriptive adjectives are organized as a single superset with 7 sets. Most of these sets contain subsets. As with most other POS in SAL, adjectives are subclassified according to the syntactic relationships that they have with other words. Table 3 summarizes the different sets and subsets of the descriptive adjective type. The sets are: pre-clausal, pre-verbal, adverbial, non-adverbial, post-nominal, and prefixes.

Pre-clausal adjectives (PC) introduce *that* clauses (*It was evident/mandatory that...*; *John was certain that...*). Some adjectives in this set may also be pre-verbal if they can be followed by a verbal complement (*It is mandatory to V*; *John is good at V'ing*). There are 3 subgroupings of pre-clausal adjectives: the pure logical, the mixed logical and the non-logical. Pure logical adjectives take the logical *it* as a subject (*It is apparent that...* or *It is urgent that/to...*, with the syntactic pattern [It is ADJ that/to]). Pure logical subsets are: *urgent* type (*that/to*), and *clear* type (*that*). Mixed logical adjectives take either a normal NP subject or the logical *it* (*She is certain that/to...* or *It is certain that/to...*, with the syntactic pattern [It/NP is ADJ that/to]). Mixed logical subsets are: *certain* type (*that/to*), and *good* type (*that/to*). Non-logical adjectives take a normal NP subject, but cannot take the logical *it* (*He is happy that/to...* or *She is hopeful that...*, with the syntactic pattern [NP is ADJ that/to]). Non/logical subsets are: *happy* type (*that/to*) and *aware* type (*that*). For example, the good-type adjectives, when used as predicate adjectives, have the following characteristics: (i) they may take *that* clause complementation, (ii) they may have logical *it* subject for both *that* clause and verbal complements (*It is good that.../It is good to...*), (iii) they may have normal NP subject only for verbal complements (*John is good at V'ing*), and (iv) they may take the subjunctive (*It is good that he go*). Table 4 illustrates the different syntactic patterns for the descriptive pre-clausal good-type adjectives.

Pre-verbal adjectives (PV) introduce infinitive clauses (*They were eager to go*). Adjectives like *instrumental*, and *capable* introduce other types of verbal clauses (*They were instrumental in solving...*; *They were capable of solving...*). A full description of each type of pre-verbal adjectives can be found in the SAL Tutorial.

Adverbial adjectives (AV) (*slight*, *real*, *minimal*) are a broad adjectival class distinguished by important characteristics. One characteristic is that they denote adverbial concepts of manner, place, time, degree, etc. and always have adverb counterparts (*slightly*, *really*, *minimally*). This classification allows the system to transform phrases like *rapid oscillation* to *oscillate rapidly*, a transformation often called for in certain target languages. The adverbial adjective set is further broken down into adjective sub-

	SAL id	Mnemonic	Example
Pre-clausal	13 86 432	PCurgent	*urgent, essential, crucial, feasible, absurd, appropriate, compulsory, convenient*
	13 86 433	PCclear	*clear, apparent, arguable, evident, implicit, inevitable, ironic, obvious, pertinent*
	13 87 438	PCcertain	*certain, curious, fortunate, lucky, sure, unfortunate, unlikely, likely*
	13 87 439	PCgood	*good, excellent, horrible, nice, smart, great, odd, mad, prudent, vital, optimistic*
	13 88 442	PChappy	*happy, afraid, anxious, desperate, proud, ashamed, proud, furious, glad, grateful*
	13 88 443	PCaware	*aware, apologetic, hopeful, confident, angry, exuberant, insistent, joyful, unsure*
Pre-verbal	13 83 418	PVvalid	*valid, beneficial, common, healthy, worthwhile, useful, profitable, meaningless*
	13 84 419	PVakin	*akin, analogous, conducive, tantamount*
	13 84 420	PVeasy	*easy, difficult, catastrophic, pleasant, costly, dangerous, hard, safe, simple, tough*
	13 84 423	PVwise	*wise, brilliant, childish, negligent, polite, astute, bestial, careless, discriminatory*
	13 85 428	PVeager	*eager, able, available, competent, free, eligible, hesitant, liable, powerless, ready*
	13 85 429	PVfirst	*first, next, slow, fifth, thirtieth, second, eighteenth, seventh*
	13 85 430	PVbusy	*busy, allergic, candid, capable, diligent, effective, forceful, proficient, resourceful*
Adverbial	13 17/ 20/ 26/ 46/ 57/ 59/ 68/ 90	AV	*envious (of), fluent (in), distinct (from), famous (for), adjacent (to), familiar (with), asleep (on), joyous (about/over), reliant (on)*
	13 80 180	AVstate	*charismatic, exhaustive, irreverent, jovial, joyless, laconic, personable, graphical*
	13 80 181	AVtime	*cyclical, consecutive, daily, final, recent, former, hourly, immediate, momentary*
	13 80 182	AVloc	*above, backward, inward, nationwide, outbound, overseas, regional, skyward*
	13 80 183	AVquan	*approximate, endless, enormous, lenghty, colossal, countless, minimal, numerous*
	13 80 184	AVdegree	*acute, comprehensive, entire, substantial, intense, interminable, outstanding*
Non-Adverbial	13 81 186	NAVpred	*above-mentioned, bivalent, cordless, exploratory, far-reaching, hereditary, inferior, jobless, marbled, scholarly, untimely*
	4 81 187	NAVnpred	*Atlantic, bridal, naval, Baltic*
Post-Nominal	4 81 417	POST	*galore, asleep, unafraid, alike, alone, unsuited, aflame, alive, indebted, resistant, aground, ajar, amiss, awake, awry, aground*
Prefixes	13 82 100	PRE	*aero-, anti-, omni-, bi-, non-, contra-, infra-, micro-, multi-, semi-, pseudo-, trans-*

Table 3: Descriptive Adjectives

Pattern	Example Sentence
It is ADJ that	It is silly that...
It is ADJ for NP that	It is *good* for the employees that...
It is ADJ to VP	It is *smart* to exercise.
It is ADJ for NP to VP	It was *silly* for them to expect...
It is ADJ of NP to VP	It was *optimistic* of them to expect...
It is ADJ V'ing	It is *smart* doing the right thing.
NP is ADJ to VP	John is *smart* to exercise.
Sub-groups	
It is ADJ to NP that	It was vital to him that...
NP is ADJ V'ing	He is smart doing the right thing...
NP is ADJ at V'ing	She is good at teaching.
NP is ADJ in V'ing	He was selfish in doing this.
NP is ADJ for V'ing	Salt is good for seasoning food.
It was an ADJ NP to VP	It was a great party to attend (i.e. It was great to attend that party)

Table 4: Good-type Syntactic Patterns

sets with adverbial coloration, such as manner (*intrepid*), time (*immediate*), place (*local*), order (*previous*), and degree (*utter*). Adverbial adjectives may also govern a particular preposition (*adjacent to*, *distant from*, *indifferent about*). Either it is interchangeable with the adverb (*back*, *above*, *lower*) or it is convertible to an adverb (*immense/ immensely*, *former/formerly*, *lower/ lower*, *utter/utterly*. In addition, adverbial adjectives may (but not in all cases) function as predicate adjectives. Adverbial adjectives provide the translation system with information pertaining to stylistic transformations that may be required by the target language. For example, if an adverbial adjective is followed by the process noun *swift movement*, the NP can be transformed stylistically in the target language as *move swiftly*. Such transformations are made possible because the lexical entry for the adjective points to both a target adjectival transfer and a target adverbial transfer (called the alternate word class). On the other hand, NP's like *inner movement* are not transformable and its adjective code would tell the system not to attempt such a transformation. There are 7 subsets under the adverbial adjective set: the prep governance type, the state/manner type, the time/order type, the locative type, the quantity/measure type, and the degree (intensifier) type. A full description on each of these types of adverbial adjective can be found in the SAL Tutorial.

Non-adverbial adjectives (NAV) do not have adverbial counterparts and, therefore, cannot be converted to adverbs (*yellow*) and do not govern prepositions. The non-adverbial adjective set is distributed over two subsets: (i) predicate adjectives (adjectives which can occur in the predicate adjective position), and fit the pattern [NP is ADJ] (*The flower is yellow*); (ii) non-predicate adjectives (adjectives which cannot normally occur in the predicate adjective position, function only attributively (*Atlantic*, *bridal*, *naval*), and fit the pattern [DET ADJ N].

Post-nominal (POST) adjectives can only occur in post-nominal positions (*The lottery winner suddenly had money galore*). Pure post-nominals, like *galore*, are rare in English. These adjectives occur in the pattern [NP ADJ], and usually stand for collapsed relative clauses (for example, the phrase *The house ablaze with light* can be parsed as a shortened form of *The house that is ablaze with light*). Frequently, a pure post-nominal may govern a particular

Word Class	id	GE	FR	IT
Noun	1	28266	25910	23505
Verb	2	33855	33354	33021
Adverb (locative)	3	465	442	450
Adjective	4	21219	20749	20518
Pronoun	5	121	121	121
Adverb (manner, agency, degree)	6	2207	2167	2173
Preposition (non-locative)	11	140	140	139
Auxiliary and Modal	12	34	34	34
Preposition (locative)	13	148	148	148
Definite Article	14	194	194	189
Indefinite Article	15	66	66	65
Arithmate in Apposition	16	208	208	203
Negative	17	2	2	2
Relative and Interrogative Pronoun	18	23	23	20
Conjunction	19	160	160	160
Punctuation	20	30	30	30
Total		87138	83748	80778

Table 5: Total of Entries by POS per Target Language

preposition, and be coded for prepositional governance, as in *ablaze with*, *awash with*, *unsuited for*, *fraught with,* and *indebted to*.

Prefix adjectives appear exclusively in pre-nominal position (*anti-*, *pro-*, *omni-*). The prefix *re-* has a word-specific subset code (PREre). Since these prefixes are used with hyphenation, in the OpenLogos system, they need to be in the lexicon as distinct entities.

5. Quantitative Results

Table 5 presents the number of entries per POS and per language. There are 12 POS in SAL, divided into open and closed word classes (a total of 20). Open word classes are: nouns, adjectives, adverbs and verbs. Closed word classes are: articles, prepositions, auxiliary verbs, pronouns, interrogatives, negative particles, conjunctions and punctuation. Verbs, nouns and adjectives are clearly the most represented word classes, as they reach more than 80,000 entries.

The dictionaries are stored in self-contained XML (Extensible Markup Language) files, with the purpose of being easily addressed by small programs. The information follows the DTD (Document Type Definition) presented in Figure 1. One of the reasons for adopting the XML format is that processing XML data is now extremely facilitated by the existing efficient XML APIs (Application Programming Interfaces) that can be found for almost any programming language, such as Python, Java, etc.. It is now possible to arbitrarily process complex data structures, represented in XML format, with less than a few programming lines. Figures 2, 3, and 4 contain OpenLogos dictionary entries for English-French (verbs), English-German (nouns), and English-Italian (adjectives) in XML format.

Figure 2 illustrates 2 entries for *depart*, identified as `<Entry source=depart>`, extracted from the English-French dictionary. The first entry (target=*quitter*) is classified as TRundif (SAL: 13 98 596) which stands for the transitive undifferentiated superset. The English word follows the inflectional paradigm (PAT 01) for verbs, represented by the word *walk*, and inflects like walk, walked, walking. The French verb *quitter* follows the inflectional paradigm (PAT 03), represented by the word *parler*, with

```
<!DOCTYPE Dictionary [
<!ELEMENT Dictionary (Entry+)>
<!ELEMENT Entry (source,target)>
<!ELEMENT source (pos,morphology,sal)>
<!ELEMENT target (pos,morphology)>
<!ELEMENT pos (#PCDATA)>
<!ELEMENT morphology (inflection)>
<!ELEMENT inflection (#PCDATA)>
<!ELEMENT sal (#PCDATA)>

<!ATTLIST Dictionary source CDATA #REQUIRED>
<!ATTLIST Dictionary target CDATA #REQUIRED>

<!ATTLIST Entry source CDATA #REQUIRED>
<!ATTLIST Entry target CDATA #REQUIRED>

<!ATTLIST source word_type CDATA #REQUIRED>
<!ATTLIST source head_word CDATA #REQUIRED>
<!ATTLIST source aux CDATA #IMPLIED>
<!ATTLIST source homograph CDATA #IMPLIED>
<!ATTLIST source alternate CDATA #IMPLIED>

<!ATTLIST target word_type CDATA #REQUIRED>
<!ATTLIST target head_word CDATA #REQUIRED>
<!ATTLIST target aux CDATA #IMPLIED>
<!ATTLIST target causative CDATA #IMPLIED>
<!ATTLIST target reflexive CDATA #IMPLIED>
<!ATTLIST target aspectual CDATA #IMPLIED>

<!ATTLIST morphology gen_id CDATA #IMPLIED>
<!ATTLIST morphology num_id CDATA #IMPLIED>
<!ATTLIST morphology number CDATA #IMPLIED>
<!ATTLIST morphology gender CDATA #IMPLIED>

<!ATTLIST inflection id CDATA #IMPLIED>
<!ATTLIST inflection gender_code CDATA #IMPLIED>
<!ATTLIST inflection example CDATA #IMPLIED>
<!ATTLIST inflection description CDATA #IMPLIED>

<!ATTLIST sal code CDATA #REQUIRED>
<!ATTLIST sal superset CDATA #IMPLIED>
<!ATTLIST sal set CDATA #IMPLIED>
<!ATTLIST sal subset CDATA #IMPLIED>
<!ATTLIST sal mnemonic CDATA #IMPLIED>
<!ATTLIST sal description CDATA #IMPLIED>

<!ATTLIST pos wclass CDATA #REQUIRED>
<!ATTLIST pos description CDATA #REQUIRED>
]>
```

Figure 1: Document Type Definition (DTD)

the description regular ending in *-er*. The second entry (target=*partir*) is classified as INMOinto-type (SAL: 10 24 596) representing the motional intransitive superset. The French verb *partir* follows the inflectional paradigm (PAT 12) for irregular verbs ending in *-ir*, with shortened stem.

Figure 3 illustrates 3 entries for *crocodile*, identified as `<Entry source=crocodile>`, extracted from the English-German dictionary. The first 2 entries (target=*Ziehbank* and target=*Krokodil*) are classified as COmach (SAL: 3 35 750), representing a *concrete* noun, *machine*, *agentive*. The third entry, corresponding to target=*Krokodil*, is classified as ANreptile (SAL: 5 51 123), which stands for *animate*, *reptile*. The source word *crocodile* follows the inflectional paradigm (PAT 16) for nouns, which is represented by the word *book*, a masculine singular noun that inflects in number by adding an *-s* to the lemma. The German transfer *Ziehbank* is a feminine singular noun that follows the inflectional paradigm (PAT 57) for nouns, which is represented by the word *Hand*, whose plural is *Hände*. *Krokodil* is classified with the neuter gender (gen_id=3) and singular number (num_id=1) in entries 2 and 3. The inflectional paradigm for *Krokodil* is represented by the word *Ziel* (PAT 68), with the description -s/-e like Ziel/Ziele.

```
<Entry source="depart" target="quitter">
  <source head_word="1" homograph="no" word_type="01">
    <pos description="Verb" wclass="02"/>
    <morphology>
      <inflection description="like walk, walked, walking" example="walk" id="1"/>
    </morphology>
    <sal code="13,98,596" description="create, etc." mnemonic="generictransitive4" set="other98"/>
  </source>
  <target aux="1" head_word="1" word_type="01">
    <pos description="Verb" wclass="02"/>
    <morphology>
      <inflection description="regular ending in -er: parler" example="parler" id="3"/>
    </morphology>
  </target>
</Entry>
<Entry source="depart" target="partir">
  <source head_word="1" homograph="no" word_type="01">
    <pos description="Verb" wclass="02"/>
    <morphology>
      <inflection description="like walk, walked, walking" example="walk" id="1"/>
    </morphology>
    <sal code="10,24,596" description="from = away from, off of, out of" set="governsawayfrom"/>
  </source>
  <target aux="2" head_word="1" word_type="01">
    <pos description="Verb" wclass="02"/>
    <morphology>
      <inflection description="Irreg. in -ir with shortened stem ..." example="partir" id="12"/>
    </morphology>
  </target>
</Entry>
```

Figure 2: Example for the entry *depart* extracted from the English-French dictionary

```
<Entry source="crocodile" target="Ziehbank">
  <source head_word="1" homograph="no" word_type="01">
    <pos description="Noun" wclass="01"/>
    <morphology num_id="1" number="singular">
      <inflection description="like book, books" example="book" id="16"/>
    </morphology>
    <sal code="3,35,750" mnemonic="machine" set="agentive" subset="machine" superset="concrete"/>
  </source>
  <target head_word="1" word_type="01">
    <pos description="Noun" wclass="01"/>
    <morphology gen_id="2" gender="feminine" num_id="1" number="singular">
      <inflection description="-/-" like Hand/Hände" example="Hand" gender_code="2" id="57"/>
    </morphology>
  </target>
</Entry>
<Entry source="crocodile" target="Krokodil">
  <source head_word="1" homograph="no" word_type="01">
    <pos description="Noun" wclass="01"/>
    <morphology num_id="1" number="singular">
      <inflection description="like book, books" example="book" id="16"/>
    </morphology>
    <sal code="3,35,750" mnemonic="machine" set="agentive" subset="machine" superset="concrete"/>
  </source>
  <target head_word="1" word_type="01">
    <pos description="Noun" wclass="01"/>
    <morphology gen_id="3" num_id="1" number="singular">
      <inflection description="-s/-e like Ziel/Ziele" example="Ziel" gender_code="3" id="68"/>
    </morphology>
  </target>
</Entry>
<Entry source="crocodile" target="Krokodil">
  <source head_word="1" homograph="no" word_type="01">
    <pos description="Noun" wclass="01"/>
    <morphology num_id="1" number="singular">
      <inflection description="like book, books" example="book" id="16"/>
    </morphology>
    <sal code="5,51,123" mnemonic="reptile" set="animate" subset="reptile" superset="animate"/>
  </source>
  <target head_word="1" word_type="01">
    <pos description="Noun" wclass="01"/>
    <morphology gen_id="3" num_id="1" number="singular">
      <inflection description="-s/-e like Ziel/Ziele" example="Ziel" gender_code="3" id="68"/>
    </morphology>
  </target>
</Entry>
```

Figure 3: Example for the entry *crocodile* extracted from the English-German dictionary

```
<Entry source="happy" target="felice">
  <source head_word="1" homograph="no" word_type="01">
    <pos description="Adjective" wclass="04"/>
    <morphology>
      <inflection description="like early, earlier, earliest" example="early" id="33"/>
    </morphology>
    <sal code="13,88,442" set="preclausalnonlog" subset="happy" superset="descriptive"
        description="non-logical [he]; that/to" />
  </source>
  <target head_word="1" word_type="01">
    <pos description="Adjective" wclass="04"/>
    <morphology>
      <inflection description="adj with 1 sing in -e (m and f), 1 pl in -i (m and f), sup in -issimo" id="46"/>
    </morphology>
  </target>
</Entry>
```

Figure 4: Example for the entry *happy* extracted from the English-Italian dictionary

Figure 4 illustrates the adjectival entry *happy* (word class 04), identified as `<Entry source=happy>`, extracted from the English-Italian dictionary, whose target is the word *felice*. The source word is classified as a descriptive pre-clausal (non-logical) happy-type adjective (SAL: 13 88 442), which is used as a predicate adjective, as described in Section 4.3.. The inflectional paradigm of *happy* is (PAT 33) following the morphology of the adjective *early*. The comparative and superlative of *happy* are formed by the removal of the *-y* and the addition of the suffixes *-ier* and *-iest*. Its Italian transfer *felice* follows the inflectional paradigm (PAT 46) for adjectives ending in *-e* in the singular, and in *-i* in the plural, both for masculine and feminine forms. The superlative form ends in *-issimo (issima, issimos, issimas*).

6. Conclusions and Future Work

The main goal of this paper was to present 3 sets of resources for machine translation, which can be freely used for research purposes. These lexical resources contain semantico-syntactic knowledge concerning the conceptual formalization of things, ideas, relationships, dispositions, conditions, processes, etc., valuable for machine translation and other natural language processing applications. The resources are in XML format for easier processing. In the future, we will make available the bilingual dictionaries for English-Portuguese, English-Spanish and German-English.

Acknowledgements

Anabela Barreiro was funded by FCT (post-doctoral grant SFRH / BPD / 91446 / 2012). This research was also supported by FCT, under project PEst-OE/EEI/LA0021/2013.

7. References

B. Vandecapelle A. Raw and F. Van Eynde. 1988. Eurotra: an overview. *Interface. Journal of Applied Linguistics*, 1(3):5–32.

M-H. Antoni-Lay, G. Francopoulo, and L. Zaysser. 1994. A generic model for reusable lexicons: The GENELEX project. *Literary and Linguistic Computing*, 9(1):47–54.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anabela Barreiro, Bernard Scott, Walter Kasper, and Bernd Kiefer. 2011. Openlogos rule-based machine translation: Philosophy, model, resources and customization. *Machine Translation*, 25(2):107–126.

Nicoletta Calzolari, Alessandro Lenci, Francesca Bertagna, and Antonio Zampolli. 2002. Broadening the scope of the eagles/isle lexical standardization initiative. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization - Volume 12*, COLING '02, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Rebecca Green, Bonnie J. Dorr, and Philip Resnik. 2004. Inducing frame semantic verb classes from wordnet and ldoce. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 375–382, Barcelona, Spain, July.

Alessandro Lenci, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. 2000. Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Eugenio Picchi, Carol Peters, and Nicoletta Calzolari. 1988. Implementing a Bilingual Lexical Database System. In T. Magay and J. Zigány, editors, *Proceedings of the 3rd EURALEX International Congress*, pages 317–329, Budapest, Hungary, sep. Akadémiai Kiadó.

Paul Procter, editor. 1978. *Longman Dictionary of Contemporary English*. Longman Group Ltd., Essex, UK.

Bernard (Bud) Scott. 2003. The logos model: An historical perspective. *Machine Translation*, 18(1):1–72.

Marta Villegas, Núria Bel, Alessandro Lenci, Nicoletta Calzolari, Nilda Ruimy, Antonio Zampolli, Teresa Sadurní, and Joan Soler. 2000. Multilingual linguistic resources: From monolingual lexicons to bilingual interrelated lexicons. In *LREC*. European Language Resources Association.

P. Vossen. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.