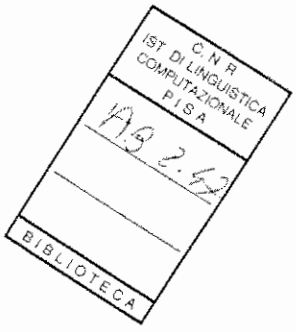
AN OVERVIEW OF WORK ON SEMANTIC TAXONOMIES

Amsterdam, Cambridge and Pisa



April 1991

ESPRIT BRA 3030 ACQUILEX WP No.29 Bis

AN OVERVIEW OF WORK ON SEMANTIC TAXONOMIES

Part 1: Cambridge and Amsterdam

Ann Copestake, Antonio Sanfilippo, Piek Vossen

Part 2: Pisa

A. Alonge, N. Calzolari, J. Hagman, E. Marinai, S. Montemagni, C. Peters, E. Picchi, A. Roventini, A. Spanu, A. Zampolli

An overview of work on semantic taxonomies

Ann Copestake, Antonio Sanfilippo and Piek Vossen

April 1991

ESPRIT BRA-3030 ACQUILEX Deliverable No. 2.3.8
Semantic taxonomies

This document is an overview of the current status of the work on taxonomies in the ACQUILEX project. It has been written as part of the ACQUILEX Deliverable No. 2.3.8 (the supply of semantic taxonomies). It summarises the work on the extraction and comparison of taxonomies and their use in constructing the lexical knowledge base (the LKB).

As far as extraction is concerned, in this document we concentrate on the work on LDOCE. The distribution of taxonomies mainly involves Cambridge providing taxonomies derived from LDOCE to the other sites, since most of the multi-lingual work is being carried out at Amsterdam, Pisa and Barcelona. LDOCE English is, in effect, being used as a core-language for the multi-lingual work on the project for several reasons:

1. The testbed will demonstrate generation of translations from English to the other languages.

2. Limited availability of dictionaries (Italian-English and Dutch-English being the only bilingual ones.)

3. LDOCE uses a restricted vocabulary and the other dictionaries do not. Taxonomies extracted from LDOCE therefore tend to have a relatively simple structure; this makes them appropriate as a general target onto which other taxonomies can be mapped.

4. Transfer of information will most often go from LDOCE to other dictionaries (and not vice versa) because of the greater detail of grammatical coding etc in LDOCE.

Work on taxonomy extraction has, of course, proceeded at other sites; see, for example, Alonge(1991), Vossen and Serail (1990), Vossen (1991b) and Rodriguez et al (1990).

1 Creation of noun taxonomies.

The software described in Copestake(1990a,b), for the semi-automatic creation of disambiguated taxonomies, has been used to build LDOCE IS_A noun taxonomies, starting from senses of "animal", "plant", "person", "man", "woman", "substance", "instrument" and so on, covering about 60% of the concrete nouns in LDOCE (7,500 word senses). This made use of the work on parsing LDOCE definitions carried out by Vossen (1990a). These taxonomies were built at a rate of 500-1000 word senses per hour (depending on the degree of interaction needed, determined largely by the potential ambiguity in the genus term). Errors are normally localised and thus easy to detect. The rate of failure of the heuristics depends on the taxonomy being built, but seems acceptable even in the worst cases. (Unedited taxonomies derived from the various senses of instrument, a relatively difficult example, are appended to this document.)

The taxonomy creation program is a general tool which is used in conjunction with our lexical database software (LDB) and can be customised for use on any MRD where the definitions can be parsed to produce an undisambiguated genus term. The taxonomies created can be stored and queried in the LDB in conjunction with

the dictionary for which they were derived. The basic program has been used with different heuristics for LDOCE verbs (for which the LDOCE box codes are in general unhelpful) and for work on the Spanish monolingual dictionary, VOX (Rodriguez et al 1990), although disambiguation in both these cases is more difficult. Preliminary results for LDOCE verbs suggest that because some assignments are checked by the user, reasonable results can be obtained even with less reliable heuristics, although more interaction time is required.

2 Distribution of noun taxonomies

Taxonomies are being distributed to other sites in three formats:

1. The most basic format is designed to be human readable, with taxonomy depth indicated by indentation etc. The example taxonomies appended to this document are in this format. Such a representation is useful for manual comparison of taxonomies. Taxonomies extracted from LDOCE covering persons, instruments and substances have been distributed to all sites in this format. These taxonomies include those words in the agreed subset which are straightforwardly retrievable via taxonomies of this type (see below).

2. The output from the semi-automatic taxonomy creation program can also be used to build a derived dictionary in the LDB. This enables the taxonomies to be queried in conjunction with LDOCE (at those sites which are using the LDB and which have access to LDOCE) so producing queries to retrieve, for example, all the count nouns that are in the liquid taxonomy is very straightforward. (Taxonomies stored in the LDB can also be queried without the parent dictionary, but obviously this is not as useful.)

3. The main use of the extracted taxonomies is expected to be in structuring the LKB. The prototype version of the LKB software was distributed to all partners at the 18 month workshop. An example LKB lexicon containing entries automatically derived from the "drink" taxonomy, with the taxonomic links converted into default inheritance relationships, was also distributed at this point.

Rather than rigidly defining a list of words as a vocabulary subset on which to work jointly, we have identified some (semantic) classes for which we will attempt to extract word senses and represent them as completely as is feasible. The currently agreed classes for nouns are:

1. Foods and drinks (possibly liquids in general). "Food" and "drink" (in the appropriate senses) form part of the LDOCE substance taxonomy. The majority of word senses denoting foods and drinks can be found by extracting the taxonomies but other word senses will have to be retrieved by analysis of the differentia.

2. Instruments used in cooking (similar comments apply).

3. Representation nouns (eg "book", "picture"; taxonomies are of less use in extracting these from LDOCE).

4. Professions and habitual occupations. Work on this set will be a combination of work on noun taxonomies and work on derivational morphology of verbs, since many examples are derived forms (eg "bake", "baker").

These classes may be changed if necessary and other classes may be added if time permits (for example, places of work).

3 Use of taxonomies

As described in Copestake(1990a) the taxonomies are used in the LKB to provide a hierarchy for default inheritance. The default inheritance mechanism currently being used is described in Copestake et al (1991). The use of the LKB software makes it possible to focus on solutions to the problems identified in the representation of taxonomies in general; for example, circularities, coordination of genus terms, the structure of the top level of the taxonomies, non-hyponymy relations, variations in classification (see Vossen 1990b, 1991a). Preliminary solutions to some of these problems are emerging — for example, the use of types to provide the "top-level" structure in the LKB, the use of multiple inheritance to deal with variations in classification. We can also formalise some aspects of coordination and of non-hyponomy relations in the LKB (Copestake et al 1991). The use of types to define appropriate features provides a way of defining the "relativised qualia structure" templates for noun taxonomy attributes (Calzolari 1991). We therefore expect that much cross-linguistic work from now on will involve exchanging information in the LKB format.

Taxonomies are also useful in work on sense-extension/regular polysemy and derivational morphology (see Vossen 1990b and Briscoe and Copestake 1991). We can use the extracted taxonomies to help define classes of words which undergo regular sense extension (which may or may not be lexicalised). For example words with a primary sense denoting an animal may also have senses denoting the meat of the animal, the fur of the animal, or a metaphorical usage denoting a human. By identifying and representing these sense extensions we can augment the lexicalised senses; we can also attempt to quantify the extent to which polysemy is regular (for a given class of words). We expect similar techniques to be useful for derivations.

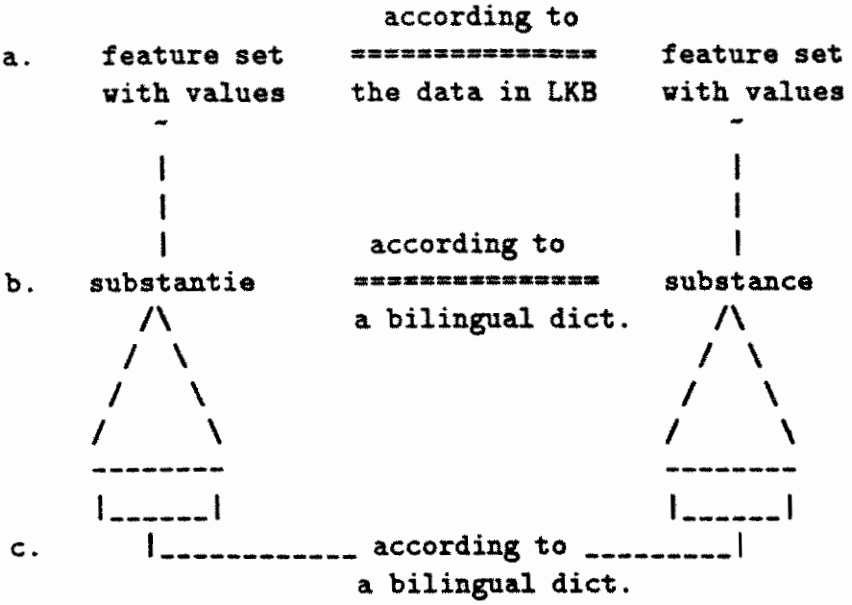
We intend to use taxonomies as a basis for comparison and merging of information cross linguistically. If we take for instance the Dutch word "substantie" and the English word "substance" then there are three ways of comparing the data, schematically represented as shown in Figure 1.

1. Determine whether two word senses could be equivalent according to the information stored in the LKB.

2. Determine equivalence according to a bilingual dictionary.

3. See whether two words are used as a genus term in the definitions of a comparable group of words (using a bilingual dictionary to compare daughters).

We intend to examine all three approaches (and possibly combined approaches) on this project (although for the Spanish-English work only the first is feasible because Barcelona have no access to a bilingual dictionary). Vossen(1991b) discusses taxonomy comparison in detail.



4 Verb taxonomies

In addition to the work on nouns described above, we have started to derive taxonomies for verbs. We intend to use taxonomic information on verbs in conjunction with other kinds of information available through LDB queries (e.g. subcategorization) to provide an integrated characterization of syntactic and semantic properties of verbs.

In Cambridge, a considerable amount of work has already been done on deriving syntactic (and to a lesser extent semantic) information using the grammar codes given in LDOCE. With respect to verbs, for example, we are currently in a position to provide comprehensive information on complementation patterns and semantic arity. What remains to be done in order to attain a sufficiently rich lexical representation for verbs is to encode semantic information regarding eventuality types and argument roles. Our current pursuit with respect to eventuality types is to encode aktionsart information in a Vendlerian fashion (e.g. distinguishing among states, processes, accomplishments and achievements). Further distinctions might include eventuality types involved in "stage" and "individual" level predications (e.g. *be naked* vs. *be intelligent*, (Carlson 1977)). This orientation parallels the program of work which Pisa has undertaken (Alonge 1991), and that other sites have generally agreed upon.

With thematic information, the situation is more complicated since reliable role classifications are not easily defined. Consequently, we are now focussing on selectional restrictions, and will try later to "distill" specific types of "thematic" roles from those (if possible). Semantic properties which are directly involved in the determination of thematic roles — e.g. sentiency, volitionality, affectedness, movement, change of state or position, etc. — are inferred from specific genus terms (e.g. cause, make, do, change, become, move, and so on) and generalized over classes of predicates via taxonomy formation. Having established which properties can be attributed to which argument slot for each choice of predicate, the attempt will then be

made to derive thematic roles by intersecting sets of those properties. (Should this later step prove infeasible, we would then still have information regarding selectional restrictions on argument roles.)

The inference of semantic properties of verbs concerning event type and selectional restrictions is executed by forming chains through verb definitions — where each link of the chain is established by the genus term of definitions. These chains are derived from the machine readable dictionaries mounted in our LDB system using the taxonomy building software developed by Copestake (1990a,b). The taxomonies are derived in a top-down fashion starting from word senses of verbs such as *cause, move, have, be, become, feel* etc. which have been traditionally used as metalinguistic predicates in the semantic calculus of word meaning (Dowty 1979, Dowty 1988). The sample taxonomy included in the appendix provide an illustrative example of this approach. Given the time resources available, we have decided to concentrate on two specific classes of verbs: psychological and motion verbs. As of now, our work on verb taxonomies has essentially been on psychological verbs (Condoravdi and Sanfilippo 1990). (This fact is reflected in the taxonomy below where chains leading to verb entries other than psychological predicates have been generally blocked.)

References

Alonge A(1991) *Extraction of information on Aksionsart from verb definitions in machine-readable dictionaries,* ACQUILEX WP no ??, to be presented at the specialised conference on Natural language processing and its Applications within the 11th International Workshop on Expert systems and their applications, Avignon

Briscoe E J and Copestake A(1991, forthcoming) 'Lexical Operations in a Unification-based Framework', *Proceedings of the SIGLEX Workshop on Knowledge Representation and Lexical Semantics,* Berkeley, California

Calzolari N(1991) *Representation of semantic information in a lexical knowledge base,* ACQUILEX WP no ??, presented at the ACQUILEX 12 month review, Cambridge

Carlson G(1977) *Reference to Kinds in English,* PhD Thesis, University of Massachusetts

Condoravdi C and Sanfilippo A(1990) *Notes on Psychological Predicates,* ACQUILEX working paper No 007

Copestake A(1990a) 'An Approach to Building the Hierarchical Element of a Lexical Knowledge Base from a MRD', *Proceedings of the International Workshop on Inheritance in Natural Language Processing,* Institute for Language Technology and Artificial Intelligence, Tilburg University, The Netherlands, pp.19–29

Copestake A A(1990b) *A system for building disambiguated taxonomies,* ACQUILEX working paper No 12

Copestake A A, de Paiva V C V, Sanfilippo A and Briscoe E J(1991) *Functionality of the LKB,* ACQUILEX deliverable 2.3.7

Dowty D(1979) *Word Meaning and Montague Grammar,* Reidel, Dordrecht

Dowty D(1987) *Thematic Proto Roles, Subject Selection, and Lexical Semantic Defaults,* LSA Colloquium paper

Rodriguez H, Marti T, Verdejo F(1990) *An interactive environoment for the extraction and management of taxonomies for MRDs,* Presented at the ACQUILEX 12 month review, Cambridge

Vossen P(1990a) *A Parser-Grammar for the Meaning Descriptions of LDOCE,* Links Project Technical Report 300-169-007, Amsterdam University

Vossen P(1990b) *The end of the chain: where does decomposition of lexical knowledge lead us eventually?,* ACQUILEX working paper No 010 (also to appear in the proceedings of the 4th conference on Functional Grammar, June 1990 Kopenhagen)

Vossen P(1991a) *Extracting taxonomies from dictionary definitions,* Paper presented at the ACQUILEX workshop on default inheritance, Cambridge

Vossen P(1991b) *Comparing noun taxonomies cross linguistically,* ACQUILEX working paper No ??

Vossen P and Serail I(1990) *Word-Devil: A taxonomy-browser for decomposition via the lexicon,* ACQUILEX working paper No 009

A Example noun taxonomy derived from LDOCE

```
;;; Unedited
;;; Taxonomy for instrument 0 (1)

instrument 0 (1)
  altimeter 0 (0)
  ammeter 0 (0)
  amplifier 0 (0)
    amp 0 (2)
  aneroid barometer 0 (0)
  astrolabe 0 (0)
  audiometer 0 (0)
  auger 0 (0)
  automatic pilot 0 (0)
  balance 1 (1)
  barometer 0 (1)
    barograph 0 (0)
    weatherglass 0 (0)
  bellows 0 (1)
  brush 2 (1)
    besom 0 (0)
    broom 0 (2)
    brushwood 0 (0)
    hairbrush 0 (0)
    nailbrush 0 (0)
    paintbrush 0 (0)
    scrubbing brush 0 (0)
    toothbrush 0 (0)
    whisk 1 (2)
  card 2 (0) BLOCKED
  cathode ray tube 0 (0)
  chronograph 0 (0)
  clock 1 (1)
    alarm clock 0 (0)
    cuckoo clock 0 (0)
    grandfather clock 0 (0)
    watch 2 (1)
      fob watch 0 (0)
      hunter 0 (4)
      stopwatch 0 (0)
      ticker 0 (3)
      wristwatch 0 (0)
  clothes peg 0 (0)
  compass 2 (3)
  compass 2 (2)
  compass 2 (1)
  curling iron 0 (0)
  cutter 0 (3)
  detector 0 (0)
  device 0 (1)
  die 2 (2)
  dipswitch 0 (0)
  dividers 0 (0)
  double bass 0 (0)
  extinguisher 0 (0)
  fan 1 (0)
    punkah 0 (0)
  flyswatter 0 (0)
  forceps 0 (0)
  fork 1 (1)
    carving fork 0 (0)
    hayfork 0 (0)
    toasting fork 0 (0)
  gauge , "46 AmE "44 also "45 gage 1 (4)
  Geiger counter 0 (0)
  glass 0 (4)
  grapnel 0 (0)
    grappling iron 0 (0)
  grater 0 (0)
    shredder 0 (1)
  guillotine 1 (2)
  heliograph 0 (0)
  key 1 (1)
    latchkey 0 (0)
    master key 0 (0)
    passe - partout 0 (2)
    passkey 0 (1)
    passkey 0 (2)
    skeleton key 0 (0)
  knocker 0 (2)
    doorknocker 0 (0)
  laryngoscope 0 (0)
  lens 0 (2)
  level 1 (6) BLOCKED
  lie detector 0 (0)
  lighter 2 (2)
  manometer 0 (0)
  megaphone 0 (0)
    loudhailer 0 (0)
  micrometer 0 (0)
  microphone 0 (0)
    mike 0 (0)
  microscope 0 (1)
    electron microscope 0 (0)
  mileometer , "45 milometer 0 (0)
  mine detector 0 (0)
  monitor 1 (5)
  nail file 0 (0)
  pantograph 0 (0)
  parer 0 (0)
  pedometer 0 (0)
  pen 3 (1)
    ballpoint 0 (0)
      biro 0 (0)
    felt - tip pen 0 (0)
    fountain pen 0 (0)
  pessary 0 (2)
  pestle 1 (0)
  photoelectric cell 0 (2)
  photoelectric cell 0 (1)
  pick 3 (1)
  pitot tube 0 (0)
  pressure gauge 0 (0)
  probe 1 (1 2)
    space probe 0 (0)
  prod 2 (2)
  protractor 0 (0)
  quadrant 0 (2)
  rack 2 (3)
    wrack 1 (0)
  racket , "45 racquet 1 (0)
  rain gauge 0 (0)
  range finder 0 (0)
  rattle 2 (2)
  razor 0 (0)
    cutthroat 0 (2)
  rectifier 0 (2)
  reflecting telescope 0 (0)
  refracting telescope 0 (0)
  regulator 0 (0)
  safety razor 0 (0)
  salinometer 0 (0)
  scanner 0 (0)
  scope 2 (0)
  scuba 0 (0)
  seismograph 0 (0)
  sextant 0 (0)
  shuttle 1 (1)
  sitar 0 (0)
  slide rule 0 (0)
  speedometer 0 (0)
  spyglass 0 (0)
  squeezer 0 (0)
  stapler 0 (0)
  starter 0 (4)
  strainer 0 (0)
  stylus 0 (2)
  stylus 0 (1)
    style 1 (8) BLOCKED
  telemeter 0 (0)
  telescope 1 (0)
  tenor 0 (2)
  theodolite 0 (0)
  thermometer 0 (0)
    clinical thermometer 0 (0)
  tongs 0 (0)
  transmitter 0 (2)
  treble 1 (2)
  trephine 1 (0)
  tuba 0 (0)
  vibrator 0 (0)
  wind gauge 0 (0)

;;; Taxonomy for instrument 0 (2)

instrument 0 (2)
  accordion 0 (0)
    piano accordion 0 (0)
  alto 1 (3)
  bagpipes 0 (0)
    pipes 0 (0)
  balalaika 0 (0)
  banjo 0 (0)
  barrel organ 0 (0)
  basset horn 0 (0)
  bassoon 0 (0)
  bouzouki 0 (0)
  bugle 0 (0)
  call 2 (3) BLOCKED
  castanets 0 (0)
  cello 0 (1)
    violoncello 0 (0)
  chime 1 (4)
  clarinet 0 (0)
  clarion 0 (1)
  clavichord 0 (0)
  concertina 1 (0)
  cor anglais 0 (0)
  cornet 0 (1)
  drum 1 (1)
```

```
    bongo 0 (0)
    kettledrum 0 (0)
      timpani 0 (0)
    snare drum 0 (0)
    tabor 0 (0)
    tambour 0 (2)
    tom - tom 0 (1)
    tom - tom 0 (2)
  dulcimer 0 (0)
  euphonium 0 (0)
  flute 1 (0)
  French horn 0 (0)
  glockenspiel 0 (0)
  guitar 0 (2)
  guitar 0 (1)
  harmonium 0 (0)
  harp 0 (0)
  harpsichord 0 (0)
    spinet 0 (1)
  horn 0 (6)
  Jew's harp 0 (0)
  kazoo 0 (0)
  lute 1 (0)
  mandolin 0 (0)
  marimba 0 (0)
  metronome 0 (0)
  mouthorgan 0 (0)
    harmonica 0 (0)
  nose flute 0 (0)
  oboe 0 (0)
    hautbois , "45 - boy 0 (0)
  ocarina 0 (0)
  organ 0 (5)
    gland 0 (0)
    gonad 0 (0)
    liver 1 (1)
    spleen 0 (1)
  organ 0 (4)
  panpipes 0 (0)
  percussion 0 (2)
  piano 2 (0)
    concert grand 0 (0)
    grand piano 0 (0)
    player piano 0 (0)
    spinet 0 (2)
    upright piano 0 (0)
  piccolo 0 (0)
  psaltery 0 (0)
  record player 0 (0)
  recorder 0 (1)
  reed instrument 0 (0)
  rheostat 0 (0)
  saxophone 0 (0)
  soprano 1 (2)
  sousaphone 0 (0)
  stethoscope 0 (0)
  stringed instrument 0 (0)
  triangle 0 (3)
  trombone 0 (0)
    sackbut 0 (0)
  trumpet 1 (1)
  ukulele 0 (0)
  vibraphone 0 (0)
    vibes 0 (1)
  viola 1 (0)
  violin 0 (1)
  virginal 2 (0)
    virginals 0 (0)
  whistle 1 (1)
  wind instrument 0 (0)

;;; Taxonomy for instrument 0 (3)

instrument 0 (3)
  bellows 0 (2)
  bypass 1 (2)
  distributor 0 (2)
  grease gun 0 (0)
  tape recorder 0 (0)
  thumbscrew 0 (0)
  tuning fork 0 (0)
```

B Example verb taxonomy derived from LDOCE

```
;;; Taxonomy for cause 2 (0) — UNEDITED
;;; BLOCKED means that a decision has been taken to curtail top-
down
;;; derivation of the taxonomy at that point

cause 2 (0)
  abash 0 (0)
  abort 0 (2)
  accelerate 0 (2)
  accelerate 0 (1)
  acclimatise 0 (2)
  acclimatise 0 (1)
  acquit 0 (2)
  activate 0 (3)
  activate 0 (2)
  activate 0 (1)
  actuate 0 (0)
  addict 1 (0)
  addle 0 (1)
  address 1 (4) BLOCKED
  advance 1 (5)
  advance 1 (2)
  affect 2 (2)
  affect 2 (1)
  afflict 0 (0)
  age 2 (2) BLOCKED
  age 2 (1) BLOCKED
  aggregate 2 (1)
  agitate 0 (2)
  ail 0 (2)
  air 2 (2) BLOCKED
  alter 0 (1)
  alternate 2 (0)
  amalgamate 0 (2)
  ameliorate 0 (0)
  amend 0 (1)
  americanize , "45 - ise 0 (0)
  amuse 0 (2)
    disport 0 (0)
    divert 0 (3)
    entertain 0 (2)
    play 2 (2) BLOCKED
    play 2 (3) BLOCKED
    slum 2 (1)
  anchor 2 (3)
  anglicize , "45 - ise 0 (0)
  animate 2 (3)
  annoy 0 (0)
    aggravate 0 (2)
    chevy , "45 chevvy 0 (0)
    chivy , "45 chivvy 0 (0)
    get 0 (15) BLOCKED
    irk 0 (0)
    molest 0 (1)
    molest 0 (2)
    nag 2 (2)
    nark 3 (1)
    needle 2 (2)
    nettle 2 (0)
    niggle 0 (2)
    peeve 0 (0)
    pester 0 (0)
    rile 0 (0)
    tantalise , "45 - lise 0 (0)
    torment 2 (2)
  annul 0 (0)
  antagonize 0 (0)
  apply 0 (5)
  apply 0 (4)
  arouse 0 (2)
  arouse 0 (1)
  asphyxiate 0 (0)
  associate 1 (1)
  atrophy 2 (0)
  attenuate 0 (0)
  attract 0 (1)
    catch 1 (9) BLOCKED
    fetch 0 (3)
  audition 2 (1)
  augment 0 (0)
  awake 1 (2)
  awake 1 (1) BLOCKED
  back 4 (1) BLOCKED
  bake 0 (2) BLOCKED
  bake 0 (1) BLOCKED
  balance 2 (3) BLOCKED
  balance 2 (2) BLOCKED
  bang 1 (2)
  batter 1 (2)
  bawl 0 (2)
  beguile 0 (2)
  belittle 0 (0)
  bend 1 (2) BLOCKED
  bend 1 (1) BLOCKED
  bestir 0 (0)
  betake 0 (0)
  bethink 0 (0)
  better 5 (1)
  bias 2 (0)
  bind 1 (9) BLOCKED
  bind 1 (7) BLOCKED
  blacken 0 (1)
  blast 2 (3)
  bleach 1 (0)
  blend 1 (1)
```

AN OVERVIEW OF WORK ON SEMANTIC TAXONOMIES IN PISA

Pisa Group: A. Alonge, N. Calzolari, J. Hagman, E. Marinai, S. Montemagni, C. Peters, E. Picchi, A. Roventini, A. Spanu, A. Zampolli

Istituto di Linguistica Computazionale, CNR, Pisa, Italy
Dipartimento di Linguistica, Università di Pisa, Pisa, Italy

Pisa, April, 1991

List of Work Stages

1. Extraction of Taxonomies

1.1 Problems
1.2 Particular Types of Genus Terms
1.3 Quantitative Data

2. Insertion of Taxonomy Data in LDB and New Query Functions

3. Disambiguation of Genus Terms

4. Linking of Word Senses from Different Sources

5. Analysis of Taxonomies: Definition of Semantic Templates

5.1 Nouns
5.2
Verbs

6. Syntactic Analysis of Definitions

7. Rules for Semantic Patterns and their Application

8. Merging of Data from Different Sources

8.1 Taxonomies
8.2 Definition 'differentiae'

9. Cross-Linguistic Taxonomy Merging and Role of the Bilingual Dictionary

10. Insertion of Formalized Semantic Data in the LDB

1. EXTRACTION OF TAXONOMIES

Work is progressing on both Italian monolingual dictionaries (the Italian Machine Dictionary - DMI - which is mainly based on the Zingarelli Italian Dictionary, and 'Il Nuovo Dizionario Italiano Garzanti'); our strategy is to extract taxonomic information over the entire dictionaries.

In the DMI, taxonomy data has been extracted for all the noun, verb and adjective definitions (about 180,000 definitions).

In Garzanti, data has been extracted from all the noun definitions (approximately 40,000 definitions); work on the verbs has almost been completed.

The methodology adopted for identifying the genus terms is well known and has been discussed widely in the literature (see Byrd et al 1987; Calzolari 1988; Vossen and Serail 1990; Hagman, forthcoming(a)). It basically consists in finding the head of either the first NP or VP (or heads in the case of coordination), depending on whether a noun or a verb definition is being analyzed. Of course, depending on the structure of the definition, the procedure is not totally straightforward and we are now working on algorithms to correct errors in the identification of the genus. The genus identification procedure has been designed to be generalizable so that it can be applied to other dictionaries as we add them to our system.

1.1 Problems

In the following we discuss a number of the typical problems which arise when attempting to extract genus terms from dictionary definitions. These problems are common to all the partners in the project; we illustrate them by examples from our dictionaries and list the solutions which are currently being adopted in Pisa.

Treatment of coordinated genus terms:

We have chosen to consider them as two or more different genus terms (in our dictionaries most of the definitions of this type clearly result from a compacting of information which should belong to separate senses). However, the information that these genus terms cooccur in the same definition must be maintained because this is frequently an indication of a regular sense extension:

e.g. atto, effetto del + V (act, effect of + V)
che, chi (what, who / anything, anyone)
animale or persona (animal or person)
arte e scienza (art and science)

Examples of such coordination in our dictionaries are:

aborto - persona o cosa fatta male, imperfetta
cinematografia - l'arte e la tecnica della ripresa e della proiezione di spettacoli cinematografici

Circularity:

This problem is not tackled at this stage; the genus terms are maintained as found in the definitions. Circularity will be dealt with when the taxonomic data is transferred from the LDB to the LKB as a type-hierarchy (for top level nodes).

Empty words as Genus Terms:

e.g. 'tutto ciò che' (everything that)
'qualsiasi cosa che' (anything that)

Examples of such definitions in our dictionaries are:

alimento - quanto serve a mantenere in vita e a far crescere animali e vegetali

macchia - qualunque cosa che deturpi la purezza della coscienza

This type of information is normally indication of top or near top level terms.

Until now these definitions have not been treated, i.e. the syntactic genus is taken as it stands. They will be dealt with when inserting data in the LKB.

Single definitions formally recognized as two definitions:

When there is more than one definition for the same sense in the Garzanti dictionary, they are divided by a semi-colon; however, at times, what follows the semi-colon is to be taken as a continuation of the previous definition, in fact it lacks the genus term; in such cases the standard MRD parsing procedure recognizes two definitions.

For example:

mandarancio - frutto ottenuto dall'incrocio fra il mandarino e l'arancio amaro; grosso come un mandarino, con scorza sottile e liscia

mano - estremità del lato superiore formato dal polso, dalla palma, dal dorso e dalle cinque dita;
ha funzioni di organo prensile e tattile

This problem has not been treated so far but the fact that the genus term is missing could be used as a clue to resolve most of these cases (although this is not the only case in which the genus may be missing).

Definitions which refer to previous definitions:

Some definitions, usually containing a pronoun, refer to a previous definition;

Examples of such definitions in our dictionaries are:

catechismo - l'insieme dei principi cristiani esposti in domande e risposte;
- il libro che li contiene

cratere - cavità imbutiforme sulla sommità di un cono vulcanico, da cui esce la lava;
- ogni cavità di forma analoga

At present we do not treat these problems; they will be dealt with when analyzing the 'differentiae'.

1.2 Particular Types of Genus Terms

There are cases in which the definition does not (only) put the syntactic genus term(s) into a standard IS_A relation with the word defined, by means of the IS_A attribute. For instance, in definitions such as:

casale - gruppo di case in campagna
==> IS_A gruppo
SET_OF casa

abside - parte della chiesa
==> IS_A parte
PART_OF chiesa

'casale' is linked to 'casa' by a SET_OF relation and 'abside' is in a PART_OF relation to 'chiesa'. Thus, other kinds of semantic relations, i.e. attributes, in addition to IS_A can be extracted in parallel and will be formalized adequately in the LKB. (More examples of such relations are given below). However, in addition to the SET_OF and PART_OF relations which we have introduced, we have decided in this stage to leave both 'casale' and 'abside' still stand in an IS_A relation to 'gruppo', and 'parte' respectively, since, at the moment, our strategy is that every term which has been syntactically identified as the genus term is kept as such, using the normal IS_A relation. This goes for the 'empty' genus terms as well; thus a definition with a coordinated head such as:

proletario - che o chi vive esclusivamente del suo lavoro
==> IS_A che
IS_A chi
AGENT vivere

...keeps an IS_A relation to the syntactic but semantically empty genus terms 'che' and 'chi' but the relation AGENT(vivere) is also created, and will bear the feature [HUMAN: +].

Even emptier syntactic genus terms are those where an infinitive verb form is used with a nominal function or where the word is defined via 'essere' with an adjective. In this case, we create a 'technical' genus term to fill in the missing IS_A argument:

abbassamento - l'abbassare, l'abbassarsi
==> IS_A /situation/
VRB2NOUN abbassare

abilità - l'essere abile
==> IS_A /property/
ADJ2NOUN abile

Further examples of introduced relations or attributes are:

artista - chi e' abilissimo in qualche attività', anche manuale
==> IS_A chi
PROPHLDR abilissimo

addendo - ogni termine di una somma
==> IS_A termine
ELMNT_OF somma

fissaggio - l'atto, l'operazione del fissare
==> IS_A atto
IS_A operazione
VRB2NOUN fissare

In this way, by explicitly identifying the particular semantic relation and avoiding the loss of information on the syntactic genus term, we will be able to deal with this data when transferring the information to the LKB in the same way as the other partners.

1.3 Some Quantitative Data

Lists of the most frequently used genus terms for Garzanti and DMI are given in Appendix

2. INSERTION OF TAXONOMY DATA IN THE LDB AND NEW QUERY FUNCTIONS

The taxonomy data have been inserted in new fields in the LDB data structures and can now be interrogated in the same way as the rest of the lexical information in the entry using the LDB query system. New specific functions have been developed to query the genus terms which will permit the taxonomic chains to be followed both top-down and bottom-up. Once the genus term disambiguation stage has been completed (see 3 below) this procedure will be implemented in the system.

3. DISAMBIGUATION OF GENUS TERMS

A procedure has been developed for interactive sense disambiguation and can be applied to all the significant elements extracted from the various stages of parsing and analysis of the definitions. An interactive strategy was chosen as our dictionaries lack certain semantic information such as that contained in the LDOCE box codes. Each dictionary is disambiguated in terms of its own sense definitions.

This procedure is used to disambiguate the genus terms. It includes functions which make it possible to: correct the automatically identified genus term; memorize several taxonomic levels for the item being examined; add a normalized term or conceptual label to the dictionary genus term; add a conceptual label in order to permit cross-linguistic links. The procedure will be run over both our monolingual dictionaries in order to disambiguate all genus terms.

A complete description of the procedure will be given in Marinai and Picchi (1991).

4. LINKING OF WORD-SENSES FROM DIFFERENT SOURCES

A procedure has been studied to permit the semi-automatic linking of the definitions from our separate dictionaries in order to provide a tool which makes it easier to compare and work on data from different sources. The results can be modified interactively and saved to form part of a new merged LDB.

The procedure operates by collecting and mapping all the information for a given lemma from the two monolingual LDBs and the Italian/English bilingual LDB (based on the Collins Italian-English Concise Dictionary) onto a composite structure. Matching operations are then performed which look for identical genus terms, identical character strings in definitions and examples, equivalent subject codes and usage labels, matches between semantic indicators in the bilingual LDB and definitions in the monolinguals (and when this fails, using words identified as synonyms of the Semantic Indicators). In this way, links between the different

senses in each dictionary are formulated and a new entry in which 'equivalent' senses are mapped together and identical information is merged is proposed (for full details see Marinai et al.1991).

The procedure can be used in order to compare the taxonomy data in our two monolingual LDBs and to collect evidence on which to base proposals for taxonomy merging and the creation of 'conceptual' labels (see 8.1 below). It will also be used to pass taxonomy information, for a given word sense, from the Italian monolinguals to the translation equivalents given in the bilingual dictionary for the same word sense, thus creating links between Italian and English taxonomy data at the leaf level (see section 10 below).

5. ANALYSIS OF TAXONOMIES: DEFINITION OF SEMANTIC TEMPLATES

5.1 Nouns

The following taxonomies have been analysed manually in detail (including for some of them the equivalent LDOCE data from Cambridge):

Sostanze (Substances)

Liquidi (Liquids)

Instrumenti (Instruments)

Scienze (Sciences)

Cibi (Food): The class of 'Cibi' is differently defined in the top nodes from that of 'Food' in that, in the Italian dictionaries, a number of genus terms are used. In order to extract the whole taxonomy we have therefore to start from a set of 'top nodes' such as 'cibo, alimento, nutrimento, vivanda', which cannot be automatically derived from a single higher node. This is due a) to the fact that our dictionaries do not use a 'controlled' and 'restricted' defining vocabulary (as does LDOCE), and b) to the fact that the generic word 'cibo' is not often found in actual usage, more specific words are preferred. For this taxonomy, we shall therefore have to make a manual intervention to 'adjust' the top of the hierarchy, probably by a link not present in the dictionaries but semantically arguable from all the actual 'tops' to a node with a 'conceptual' label, 'CIBO', acting as a conveyor of the set of attributes which are common to the hierarchy.

The analysis of the above taxonomies has been carried out with the aim of identifying those patterns which are more frequently used in the 'differentia' part of the definitions, and can therefore be considered as vehicles of those types of information which lexicographers have considered relevant as defining criteria in the area considered.

Within the framework of the type or psort hierarchy presented in the definition of the formalism for the LKB (see Copestake 1990, Copestake et al 1991), for each taxonomy the nodes will be defined in terms of 'typed feature structures' corresponding to the 'meaning types' in Calzolari (1991). Our analysis of the differentiae leads to the identifcation of the attributes of these feature structures or 'meaning types'. Each attribute is instantiated in the definitions by a set of (lexically and syntactically) different patterns which, however, convey the same meaning.

For example, the following patterns:

(che è) costituito da
(che è) formato da
(che è) fatto con
a base di
avente/che ha come ingredienti

.....

can be subsumed under a single Attribute label, for example CONSTITUTED_BY. Values of this attribute will be the NP (or coordinated NPs) following the patterns.

Sets of basic attributes have thus been defined for the above taxonomies. Those defined for the taxonomy of 'SOSTANZE' have already been checked against the English taxonomy of 'SUBSTANCES', evidencing a correspondence between the type of information recorded in the different dictionaries of different languages.

5.2 Verbs

The following taxonomies have been analyzed manually in detail:

Agire (to act, to operate)
Causare (to cause)
Colpire (to hit)
Compiere (to accomplish, to make)
Correre (to run)
Diventare (to become)
Dividere (to divide)
Essere + adj. (to be ...)
Muoversi (to move)
Rendere (to make, to cause)
Sentire (to feel)
Separare (to separate)

For these taxonomies, the analysis so far has aimed at: 1) identifying the Aktionsart of the verbs in the taxonomies considered (see Alonge, 1991); 2) extracting information on selectional restrictions and thematic roles (see Calzolari 91 and Section 7 of this document); 3) associating groups of taxonomies under the same 'conceptual label', see Section 8.1 of this document. By selecting particular genus terms (i.e. 'rendere' and 'diventare') which correspond to the verbs considered as 'atomic predicates' in studies on verb semantics (see, for example, Dowty 1979) and by building their taxonomies, we have been able to single out verbs which undergo the so-called causative/inchoative alternation (see Levin 1989). This phenomenon is dealt with in Roventini and Antelmi (1991). In the future we shall attempt to identify other genus terms which can be used to recognize verbs exhibiting different diathesis alternations. Further work is now being done in order to extract information on verb subcategorization. Unfortunately, the information that we find in our dictionaries in this respect is not systematic, so that we are now in the stage of elaborating methodologies both for identifying the information available and/or performing the cross-linguistic transfer of information from LDOCE (given that in Cambridge much work has already been done in relation to syntactic information, using LDOCE existing grammar codes).

In order to make all the information which is contained in our dictionaries with respect to verbs explicit, we have tried to build attribute templates for verb classes (i.e. taxonomies) in a similar way to what has been done for nouns, i.e. identifying patterns related to attributes which often represent thematic roles. Some examples of templates are given below:

MUOVERSI (to move)

Manner:
Source:
Path:
Goal:
By_means_of (transport):

Typ. Subj.:

RENDERE (to make, to cause)

Result:
  Quality:
  Apt_to:
  More/less/equal_to:
By_means_of:

Typ. Obj. (patient):

COLPIRE (to hit)

Instrument:
Manner:
Location:
Iteratvity:

Typ. Obj. (patient):

6. SYNTACTIC ANALYSIS OF DEFINITIONS

The PLNLP Italian Grammar, conceived as the first component of a broad coverage Natural Language system, is our main tool for analyzing dictionary definitions (see Chonod et al. 1991). For our purposes, it has been integrated with a smaller component designed to handle syntactic phenomena which are specific to dictionary definitions rather than textual corpora and to disambiguate descriptions associated with constructions that would appear to be ambiguous in free text but not in the context of dictionary definitions.

The PLNLP Italian grammar analyzes Italian sentences by using syntactic information formalized in augmented phrase structure rules with a bottom-up, parallel parsing algorithm. The system that provides these facilities is PLNLP, or the Programming Language for Natural Language Processing (Heidorn, 1972, 1975).

The grammar rules have been written following the general strategy defined for the PLNLP English grammar, the so-called 'relaxed approach' aiming at accepting unrestricted input text (Jensen, 1986, 1988). This implies that a grammar conceived in such a way computes preliminary syntactic sketches that are syntactically consistent, but not necessarily semantically accurate. The parses produced contain syntactic and - whenever possible - functional information, but no semantic information or other information beyond the functional level. In Italian, in some cases, not even the functional information can be assigned purely on the basis of syntactic information; in these cases, it can only be defined after background information has been acquired and evaluated within the initial analysis.

The analysis of a sentence using only syntactic information can result in much ambiguity. We have mentioned the ambiguity inherent in the assignment of the

functional roles. The other main source of ambiguity to be considered is the attachment of modifiers (prepositional phrases, relative and other embedded clauses). The method adopted for dealing with both kinds of ambiguity is to collapse the different syntactic descriptions within the same structure, whenever possible. The solution for attachment ambiguity is to attach modifiers in a single arbitrary pattern (usually the closest possible head), but to mark other possible attachment sites so that they can be referred to for later semantic processing. When treating assignment ambiguity, we code the possible interpretations within the same structure in order to prepare it for further processing stages. This is the reason why we usually think of the resulting analysis as a 'syntactic sketch', or as an 'approximate parse'.

The organized structure resulting from this initial stage of the analysis can then be revised and sometimes disambiguated on the basis of the peculiarities of the language used within dictionary definitions. This is the case, for instance, of the attachment of prepositional phrases to conjoined genus terms (instead of to the nearest possible one, that is to the last conjoined constituent), or of the assignment of functional roles (we assume that constructions used within dictionary definitions are always unmarked). This component, still under development, should take care of the phenomena that can be considered as dictionary language specific and, at the same time, should refine the analysis produced during the first stage.

For further details on the PLNLP Italian Grammar and its use on a corpus of dictionary definitions, see Montemagni (1991).

7. RULES FOR SEMANTIC PATTERNS AND THEIR APPLICATION

The procedure for the extraction of semantic information from dictionary definitions and its consequent formalization begins with the syntactic sketch produced by the PLNLP Italian grammar. Semantic information is extracted in two stages. First, the grammar is applied to the dictionary definition to derive one or more parse trees. After the analysis has been computed, the system will apply a pattern-matching mechanism look-up to these parse trees .

The component for extracting semantic information from dictionary definitions and for formalizing the results in sets of attribute-value pairs will consist in a set of procedures written in PLNLP.

Each pattern is formally represented as a formula to be applied iteratively through the parse tree associated with a dictionary definition. Such formulas exploit the fact that every parse tree node as well as every word is represented in PLNLP as a record structure with attributes and values. Each formula corresponding to a pattern describes, in PLNLP terms, the syntactic environments within which the same semantic relation can be expressed. However, the same syntactic construction can be used to express different semantic relations. This fact led us to distinguish two types of patterns:

1) patterns describing general syntactic constructions such as
NOUN_DI_NOUN, PRON_VP / NOUN_RELCL,
independently of the particular semantic relation with which they could be associated;

2) patterns identifying semantic relations.

We refer to the first kind of patterns with the label of 'pre-pattern', given that they play the role of filter in the selection of the semantic patterns to be applied to the nodes of the parse tree. The other kind of patterns have the twofold function of describing the peculiarities of the semantic pattern and of building a formal representation of the semantic information extracted. If a record corresponding to the description given in the patterns is found in the parse tree, its head should be returned as value of the attribute corresponding to the pattern, together with other information of interest in the formalization of the semantic relation.

This component is still under development, and some alternative strategies are being tested in parallel. The final results will be described in Hagman (forthcoming(b)).

8. MERGING OF DATA FROM DIFFERENT SOURCES

8.1 Taxonomies

In order to overcome the limits of individual dictionaries (incoherences, missing data, etc.), we have also begun to merge the information coming from our two sources. By analyzing some taxonomies (both nouns and verbs) we have been able to identify groups which can be associated under the same 'conceptual' label. The main reason for doing this is that there are many groups of genus terms which are defined circularly in both sources and, furthermore, words which are found to be hyponyms of one genus term of a group in one dictionary are found in the taxonomy of another genus term (of the same group) in the other.

Work so far has been concentrated on the analysis of taxonomies for:

Nouns:
Scienza - Disciplina - Studio - Branca di
Strumento - Arnese - Attrezzo - Utensile

Verbs:
Causare - Provocare - Cagionare - Procurare - Arrecare - Produrre
Agire - Operare - Esercitare
Dividere - Separare- Disgiungere - Disunire

8.2 Definition 'differentiae'

The merging of information from different dictionary sources (and also for different languages) should be made possible by the adoption of a common definition of the top-nodes of the semantic hierarchies, in which we should aim at having:

a) common top-nodes, or 'psorts' (e.g. 'SUBSTANCE', 'FOOD', 'INSTRUMENT');

b) a common definition of their feature structures, i.e. common sets of attributes;

c) a common metalanguage to describe the nodes and their attributes.

The common feature structures or 'meaning types' will act as unifying structures for information from all the sources, see Calzolari (1991).

The same strategies will be adopted when treating verb definitions.

9. CROSS LINGUISTIC TAXONOMY MERGING AND ROLE OF THE BILINGUAL DICTIONARY

As stated in Section 8.1 work is already underway at Pisa to merge taxonomies from the two Italian monolingual dictionary sources ,grouping them under the same 'conceptual labels'. We envisage the cross-linguistic linking and/or merging at two levels: the leaf nodes; the intermediate and top level nodes. This is because, whereas at the intermediate and top levels some normalization of the data is to be expected and it should to a large extent, at least theoretically, be possible to map

corresponding feature structures, at the leaf level we are often dealing with highly language and culture specific data where it is difficult to image a totally successful feature-structure mapping as frequently the leafs will have language specific idiosyncracies.

In consideration of the central role that for various reasons LDOCE plays in the project, it appears clear that English will tend to function as a close-to-metalanguage and any merging of the Italian taxonomies at a multi-lingual level will begin with a mapping to English. Using the sense mapping procedure described in 3, the senses for lemmas from our Italian monolingual dictionaries can be linked to translation equivalents in English provided by the bilingual LDB. When linking taxonomy data, at any level, a metalanguage TAX_INDICATOR can be associated, it should then be possible to locate the translation equivalent(s) in the relevant taxonomy in LDOCE, e.g. in the 'FOOD' taxonomy, the relevant senses for the Italian lemma 'pietanza' from our dictiionaries are mapped together as follows:

DMI - vivanda servita come secondo piatta
Garzanti - la vivanda che si serve a tavola dopo la minestra
Collins - course, dish

By adding a taxonomy indicator 'FOOD' to the translation equivalents given by Collins 'course' and 'dish', we provide a sense disambiguation which permits us to link them to the correct word senses in the equivalent LDOCE taxonomy. In the same way, all the food hyponyms of pietanza extracted from our monolingual dictionaries should also link to words in the LDOCE 'FOOD' taxonomy. However, Vossen (1991) shows that this method does not always work when mapping from Dutch to English. He found that words given as translation equivalents are not always found in 'equivalent' taxonomies across the two dictionaries (Van Dale and LDOCE). In any case, when the attempt to find the translation word given by the bilingual in an 'equivalent' taxonomy extracted from another dictionary (LDOCE in our case) fails, although the translation word itself appears in the headword list of the monolingual dictionary, this may be an indication of significant differences in the structures of the taxonomies in the dictionaries being compared. However, although we think that it is thus useful to attempt links between the Italian and the LDOCE taxonomies for any word-sense, we do not feel that it will be very successful to attempt to compare complete taxonomy chains across languages (especially at the lower level), since, as is clear even from a comparison between our two monolinguals, the chains tend to be dictionary dependent, especially at the lower level. We feel that at the core and top levels a comparison of conceptual labels will be of prime relevance.

10. INSERTION OF FORMALIZED SEMANTIC DATA IN THE LDB

It is our intention to insert all the results obtained from the different syntactic and semantic parses of our dictionaries in a new kind of structured field which will be implemented in the LDB system. In this way, all the intermediate results will be available and accessible for any user of the LDB. New types of access functions will be added to the query system for this purpose.

REFERENCES

Alonge A. (1991), Extraction of Information on Aktionsart from Verb Definitions in Machine-readable Dictionaries (presented at the Conference on 'Natural Language Processing and its Applications'), 11th International Workshop on Expert Systems and their Applications, Avignon, 27 - 31 May 1991.
ESPRIT BRA-3030 ACQUILEX WP No. 031

Byrd R.J., Calzolari N., Chodorow M.S., Klavans J.L., Neff M.S., Rizk O.A. (1987), Tools and Methods for Computational Lexicology, Journal of Computational Linguistics, Vol 13, Nos. 3-4, pp.219-240.

Calzolari N. (1988), The Dictionary and the Thesaurus can be Combined, in Martha Evens (ed.), Relational Models of the Lexicon, (Studies in Natural Language Processing Series), Cambridge, Mass., Cambridge University Press, pp.75-96.

Calzolari N. (1991), Acquiring and Representing Semantic Information in a Lexical Knowledge Base, to be published in the Proceedings of the Workshop on Lexical Semantics, J. Pustejovsky, (ed.) Berkely, USA
ESPRIT BRA-3030 ACQUILEX WP No. 016

Chanod J.-P., Harriehausen B., Montemagni S. (1991), Processing Multi-Lingual Argument Structures, (presented at the Conference on 'Natural Language Processing and its Applications'), 11th International Workshop on Expert Systems and their Applications, Avignon, 27 - 31 May 1991.

Copestake A.A (1990) An Approach to Building the Hierarchical Element of a Lexical Knowledge Base from a Machine Readable Dictionary, paper presented at the International Workshop on Inheritance in Natutral Language Processing, Tilburg, The Netherlands.
ESPRIT BRA-3030 ACQUILEX WP No. 008

Copestake A.A., de Pavia V.C.V., Sanfilippo A., Briscoe E.J.(1991), Functionality of the LKB, ACQUILEX Deliverable 2.3.8

Dowty D.R. (1979), Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague PTQ", in Synthese Language Library, 7, Dordrecht, B.Reidel.

Hagman J., Common and Odd Relations in Italian Dictionaries and their Treatment in Taxonomy Building
ESPRIT BRA-3030 ACQUILEX WP, forthcoming (a).

Hagman J., Semantic Parsing of Italian Dictionary Definitions
ESPRIT BRA-3030 ACQUILEX WP, forthcoming (b).

Heidorn G.E. (1972), Natural Language Inputs to a Simulation Programming System, Ph.D. dissertation, Yale University.

Heidorn G.E. (1975), Augmented phrase structure grammars, in R.Schank and Nash-Webber (eds.), Theoretical Issues in Natural Language Processing, Association for Computational Linguistics.

Jensen K. (1986), PEG 1986: A Broad-Coverage Computational Syntax of English, Unpublished paper. IBM T.J. Watson Research Center, Yorktown Heights, N.Y..

Jensen K. (1988), Issues in parsing. in Proceedings of the Symposium on Natural Language at the Computer, Springer Verlag.

Levin B. (1989), English Verbal Diathesis, Lexicon Project Working Papers, 32, MIT, Cambridge, Mass.

Marinai E., Peters C., Picchi E. (1991) A Prototype System for the Semi-Automatic Sense Linking and Merging of Mono- and Bilingual LDBs, paper presented at ACH/ALLC 91, Tempe, USA, March 1991 and to be published in N. Ide and S. Hockey (eds.) Research in Humanities Computing, OUP.
ESPRIT BRA-3030 ACQUILEX WP No. 017

Marinai E., Picchi E., A Procedure for Interactive Sense Disambiguation,
ESPRIT BRA-3030 ACQUILEX WP 29

Montemagni S., Tailoring a Broad Coverage Grammar for the Analysis of Dictionary Definitions
ESPRIT BRA-3030 ACQUILEX WP 023

Roventini A., Antelmi D. (1990), Semantic Relationships within a Set of Verbal Entries in the Italian Lexical Database, to appear in the Proceedings of the IV Euralex Congress, Benalmádena, Malaga, 28 August - 1 September 1990.
ESPRIT BRA-3030 ACQUILEX WP No. 015

Vossen P. (1991), Comparing noun-taxonomies cross-linguistically
ESPRIT BRA-3030 ACQUILEX WP No. 014

Vossen P. Meijs W. and den Broeder M. (1989), Meaning and Structure in Dictionary Definitions, in B.Boguraev and E.J. Briscoe (eds.) Computational Lexicography for Natural Language Processing, Longman, London and New York, pp. 171-192.

Vossen P., Serail I (1990), Word-Devil: A taxonomy browser for decomposition via the lexicon.
ESPRIT BRA-3030 ACQUILEX WP No. 009

DICTIONARIES

Collins (1985), Collins Concise Italian-English, English-Italian Dictionary, Catherine E. Love (ed.), Collins Giunti, Firenze.

Garzanti (1984), Il Nuovo Dizionario Italiano Garzanti, Redazioni Grandi Opere Garzanti.

Zingarelli N. (1970), Vocabulario della Lingua Italiano, Zanichelli, Bologna.

APPENDIX

A. List of the most frequent genus terms in the Italian Machine Dictionary (DMI) - for nouns, verbs and adjectives.

B. Disambiguated taxonomies related to the "FOOD" class in DMI.

C. List of the most frequent genus terms in the Garzanti Dictionary - for nouns, verbs and adjectives.

D. Disambiguated taxonomies related to the "FOOD" class in Garzanti.

APPENDIX A

D B T - (E. Picchi) Dizionario Macchina Italiano DMI
- Lista delle Frequenze Decrescenti - RSem Tipo I

1	10665 - CHE	74	159 - FIGURA			78 - POSIZIONE
2	5685 - CHI	75	158 - DIVENIRE			78 - RAPPORTO
3	5071 - ATTO	76	155 - SISTEMA		127	77 - LAVORARE
4	2668 - EFFETTO	77	154 - ATTIVITA'			77 - ORNAMENTO
5	1522 - QUALITA'	78	153 - NUMERO			77 - TEORIA
6	1487 - RELATIVO	79	147 - OPERA			77 - VERSO
7	1466 - PERSONA	80	146 - MISURA		128	75 - PASSAGGIO
8	1194 - PARTE	81	145 - APPARTENENTE			75 - PROCEDIMENTO
9	1007 - ESSERE	82	144 - FRUTTO		129	74 - LIQUIDO
10	1002 - RENDERE		144 - LIBERARE		130	73 - CORRENTE
11	912 - MODO		144 - MAMMIFERO			73 - MATERIA
12	766 - FARE	83	143 - CIASCUNA			73 - REGIONE
13	736 - INSIEME	84	142 - COLORE			73 - SCRITTO
14	654 - LUOGO	85	140 - SPECIE			73 - USCIRE
15	647 - FAR	86	139 - FARSI		131	72 - FORZA
16	628 - CIO'	87	138 - PORTARE			72 - GENERE
17	592 - STRUMENTO		138 - PROCESSO			72 - STRATO
18	591 - METTERE		138 - TECNICA			72 - TRATTO
19	553 - PIANTA	88	134 - INFIAMMAZIONE		132	71 - SOMMA
20	471 - COMPLESSO	89	131 - ATTREZZO		133	70 - CONSIDERARE
21	446 - ABITANTE	90	130 - SERIE			70 - TIRARE
22	442 - DIVENTARE	91	128 - MANDARE		134	69 - CANTO
23	435 - COSA		128 - PASSARE			69 - CHIUDERE
24	405 - QUANTITA'		128 - PORRE			69 - DOCUMENTO
25	382 - DARE		128 - STRISCIA			69 - DURATA
26	376 - CONDIZIONE	92	127 - METTERSI			69 - GIOVANE
27	371 - ELEMENTO	93	126 - CORPO			69 - LASCIARE
28	364 - OPERAZIONE		126 - LEVARE			69 - PELLE
29	346 - AZIONE	94	125 - ORDINE			69 - RIUSCIRE
30	330 - MACCHINA	95	124 - CAPACITA'		135	68 - BEVANDA
	330 - STATO	96	122 - RACCOLTA			68 - ESPERTO
31	325 - SOSTANZA	97	120 - PERTINENTE			68 - PRIMA
32	323 - DONNA		120 - STARE			68 - TUTTO
33	322 - APPARECCHIO		120 - VENDITORE		136	67 - CADERE
34	307 - COLPO	98	116 - ATTEGGIAMENTO			67 - ROCCIA
35	296 - UNITA'		116 - FORMAZIONE		137	66 - ARMA
36	294 - UOMO	99	115 - FENOMENO			66 - FORNIRE
37	284 - PRENDERE		115 - TERRENO			66 - MANIFESTAZIONE
38	283 - OPERAIO	100	114 - FAMIGLIA			66 - RETE
39	275 - AVERE		114 - INSETTO			66 - SEDE
40	260 - UCCELLO		114 - LINEA		138	65 - BATTERE
41	254 - ARNESE	101	111 - COMPOSIZIONE			65 - DIALETTO
	254 - TITOLO	102	107 - NAVE			65 - POSTO
42	251 - SEGNO		107 - TEMPO			65 - SUONO
43	247 - TOGLIERE	103	106 - ATTINENTE			65 - TAGLIARE
44	244 - PRIVO	104	105 - MEMBRO		139	64 - COSTRUZIONE
45	242 - PICCOLA		105 - PRODOTTO			64 - DIVIDERE
46	240 - MANCANZA		105 - TENERE			64 - SCRITTORE
47	229 - NATIVO	105	103 - ESEGUIRE			64 - SENSO
	229 - PUNTO	106	102 - LAVORO		140	63 - DISPORRE
48	225 - TESSUTO		102 - PROPRIETA'			63 - ESPRESSIONE
49	221 - GRUPPO	107	99 - FACOLTA'			63 - FORNITO
50	220 - DISPOSITIVO	108	98 - LIBRO			63 - METODO
	220 - OGGETTO	109	97 - MUOVERSI			63 - PASTA
51	218 - MOVIMENTO		97 - VENIRE		141	62 - ARBUSTO
52	217 - ANDARE	110	96 - SALE			62 - AUMENTO
	217 - PROPRIO	111	95 - RAPPRESENTAZIONE			62 - DIGNITA'
53	215 - PIENO	112	94 - LOCALE			62 - INDIVIDUO
54	210 - TIPO		94 - TRATTARE		142	61 - CASA
55	208 - PRIVARE	113	93 - CAPO			61 - CAVALLO
56	205 - RIDURRE	114	92 - COMPONIMENTO			61 - PIANO
57	200 - ARTE		92 - PARLARE			61 - SOSTENITORE
	200 - SOTTOPORRE		92 - RUMORE			61 - STANZA
58	199 - FORMA	115	91 - ANIMALE		143	60 - ALTERAZIONE
	199 - PESCE	116	90 - SORTA			60 - CARTA
59	192 - STUDIOSO		90 - STRUTTURA			60 - CONTRATTO
60	191 - FATTO	117	88 - CONCERNENTE			60 - FERRO
	191 - PERDERE		88 - DIRE			60 - FUNGO
61	189 - PEZZO		88 - DISPOSIZIONE			60 - IMBARCAZIONE
62	185 - STUDIO		88 - VASO			60 - SIMILE
63	184 - SEGUACE	118	87 - PAROLA			60 - UNIRE
	184 - SPAZIO	119	86 - CIASCUNO		144	59 - ASTA
64	182 - LINGUA		86 - COPRIRE			59 - COMPIERE
	182 - SCIENZA		86 - GRADO			59 - DARSI
65	179 - DOTTRINA		86 - VINO			59 - DIRITTO
	179 - NOME	120	85 - EDIFICIO			59 - MANGIARE
	179 - VARIETA'		85 - PRODURRE		145	58 - ASPETTO
66	177 - ORGANO		85 - PROVARE			58 - ASSUMERE
67	171 - ALBERO		85 - SUPERFICIE			58 - DISFARE
68	169 - CARATTERE	121	84 - EMETTERE			58 - MASSA
	169 - SOLDATO		84 - ERBA			58 - NEGOZIO
69	168 - MALATTIA	122	83 - RAMO			58 - VOLGERE
	168 - PERIODO	123	82 - APERTURA		146	57 - PARTICOLARE
70	166 - RECIPIENTE		82 - COMPORTAMENTO			57 - RICCO
	166 - UFFICIO	124	80 - DEGNO			57 - TAGLIO
71	165 - TENDENZA	125	79 - CARATTERISTICA			57 - UNIONE
72	162 - COMPOSTO		79 - ZONA		147	56 - CERCARE
73	161 - DISCORSO	126	78 - ADDETTO			56 - DOLORE
	161 - GIOCO		78 - COLPIRE			56 - LEGNO
	161 - MONETA		78 - MEZZO			56 - MOLLUSCO

APPENDIX B

- Cibo 0 (1) 0 (2)
  - Ambrosia 0 (1) 0 (2)
  - Basoffia 0 (2)
  - Bazzofia 0 (2)
  - Becchime 0 (0)
  - Bocconcino 0(2)
    - Gnocco 0 (1)
    - Ignocco 0 (1)
  - Broda 0 (2)
  - Brodetto 0 (3)
  - Camangiare 0 (2)
  - Cena 0 (2)
  - Cibaccola 0 (1)
  - Colazione 0 (2) 0 (6)
    - Beruzzo 0 (1)
    - Refezione 0 (2)
  - Crema 0 (5)
    - Fondua 0 (0)
    - Fonduta 0 (0)
    - Zabaione 0 (1)
  - Cuccagna 0 (4)
  - Cucina 0 (5)
  - Dolce 0 (2)
    - Affricano 0 (3)
    - Africano 0 (3)
    - Babà 0 (0)
    - Biancomangiare 0 (1)
    - Bodino 0 (0)
    - Bonbon 0 (0)
    - Brioscia 0 (2)
    - Budino 0 (0)
    - Cedrata 0 (2)
    - Colomba 2 (3)
    - Confetto 0 (1)
      - Anacino 0 (2)
      - Anicino 0 (2)
      - Coriandolo 0 (2)
      - Diavolone 0 (0)
      - Pralina 0 (0)
    - Crema 0 (3) 0 (4)
      - Zabaglione 0 (1)
      - Zabaione 0 (1)
    - Croccante 0 (0)
      - Cubaita 0 (0)
    - Crostata 0 (0)
    - Diplomatico 0 (3)
    - Fiadone 0 (2)
    - Focaccia 0 (2)
      - Colombina 0 (2)
    - Frappa 0 (3)
    - Frittella 0 (1)
      - Crafen 0 (0)
      - Donzellina 0 (0)
      - Krapfen 0 (0)
      - Sgonfiotto 0 (1)
    - Gelato 0 (0)
      - Biscuit 0 (4)
      - Cassata 0 (2)
      - Cremolato 0 (0)
      - Giardinetto 0 (4)
      - Granita 0 (1)
      - Mantecato 0 (0)
      - Moretto 0 (3)
      - Pinguino 0 (3)
      - Spumone 0 (2)
      - Stracchino 0 (2)
    - Latte 0 (3)
    - Meringa 0 (2)
    - Monachina 0 (0)
    - Montebianco 0 (0)
    - Mostacciolo 0 (0)
    - Pampepato 0 (0)
    - Pandoro 0 (0)
    - Pangiallo 0 (0)
    - Panpepato 0 (0)
    - Pignoccata 0 (1) 0 (2)
    - Pistacchiata 0 (0)
    - Pudino 0 (0)
    - Sanguinaccio 0 (3)
    - Semifreddo 0 (0)
    - Sfogliatella 0 (0)
    - Torrone 0 (0)
    - Torta 2 (1)
      - Affricano 0 (3)
      - Africano 0 (3)
      - Carlotto 0 (1)
      - Carludovica 0 (1)
      - Cassata 0 (2)
      - Ciarlotta 0 (1)
      - Lattaiolo 0 (2)
      - Mantovana 0 (3)
      - Margherita 2 (2)
      - Millefoglia 0 (2)
      - Sfogliata 0 (0)
      - Tartufata 0 (0)
    - Zuccotto
  - Galanteria 0 (6)
  - Governime 0 (0)
  - Guastastomaco 0 (0)
  - Imbratto 0 (3)
  - Insalata 0 (1)
    - Minutina 0 (0)
  - Leccornia 0 (1)
  - Mangiarino 0 (0)
  - Manna 1 (1) 1 (3)
  - Marzapane 0 (2)
  - Merenda 0 (2)
  - Pappa 0 (3)
    - Pancotto 0 (0)
  - Pappo 2 (0)
  - Pastone 0 (2)
    - Cruscata 0 (1)
    - Pastume 0 (1)
  - Pastume 0 (3)
  - Pemmican 0 (1)
  - Pentola 0 (2)
  - Polenda 0 (1)
  - Polenta 0 (1)
    - Pattona 0 (1)
    - Polta 0 (0)
  - Poltiglia 0 (2)
    - Pesto 0 (1)
  - Postema 0 (3)
  - Pulenda 0 (1)
  - Putiglia 0 (2)
  - Salacca 0 (2)
  - Sfondastomaco 0 (1)
  - Stuzzichino 0 (2)
  - Tornagusto 0 (0)
  - Tosco 3 (0)
  - Tossico 0 (3)
  - Unto 2 (3)
  - Vivanda 0 (1)
    - Antipasto 0 (0)
      - Giardiniera 0 (3)
    - Aspic 0 (0)
    - Bollito 0 (1)
    - Cacimperio 0 (0)
    - Carnaggio 0 (2)
    - Cipollata 0 (1)

Cotto 0 (1)
Cuscus 0 (1) 0 (
Cuscuso 0 (1) 0 (
Fegato 0 (2)
Corata 0 (1)
Epatico 0 (0)
Fegatino 0 (
Fricassea 0 (1)
Imbrodolo 0 (0)
Interiora 0 (2)
Frattaglia 0 (
Regaglia 0 (
Intigolo 0 (2)
Bagnacauda 0 (
Brodetto 0 (
Civet 0 (0)
Finanziera 0 (
Salmì 0 (0)
Tocco 0 (0)
Kuskus 0 (1) 0 (
Macco 0 (1)
Manicaretto 0 (0)
Ammorsellato (0)
Borbottino 0 (
Guazzetto 0 (
Picchiante 0 (
Specialità 0 (
Spezieltà 0 (
Manzo 0 (3)
Marinato 0 (0)
Minestra 0 (1)
Bioscia 0 (2)
Boba 0 (1)
Bobba 0 (1)
Brodaglia 0 (
Bobbia 0 (
Broscia 0 (2)
Cavolata 0 (
Fagiolata 0 (
Favata 0 (1)
Minestrina 0 (
Pappa 0 (
Minestrone 0 (
Basoffia (1)
Bazzoffia (1)
Panata 0 (1)
Pappa 0 (2)
Pancotta (0)
Pastone 0 (
Cruscata (
Pastume (
Pattona 0 (2)
Sboba 0 (1)
Sbobba 0 (1)
Sbobbia 0 (1)
Sbroscia 0 (
Sciacquatura (3)
Semolino 0 (
Cremino (2)
Stracciatella (0)
Suppa 0 (1)

APPENDIX C

D B T - (E. Picchi) Dizionario Garzanti.

- Lista delle Frequenze Decrescenti - Genus

1	1807	chi	68	55	detto		28	suono
2	1688	"situation"		55	lettera		28	tumore
3	1481	atto		55	lingua		28	utensile
4	1111	effetto		55	seguace		28	varieta
5	907	"property"	69	54	apertura		28	vaso
6	801	parte		54	prodotto	92	27	abito
7	665	insieme		54	soldato		27	cavallo
8	414	persona	70	49	attrezzo		27	coma
9	402	abitante		49	composizione		27	confusione
10	281	complesso	71	48	carattere		27	danza
11	278	luogo	72	47	procedimento		27	disegno
12	272	cosa		47	proprieta'		27	legno
13	265	cio'	73	46	corpo		27	medico
14	264	strumento		46	nave		27	membrana
15	255	che		46	usato		27	osso
16	253	pianta	74	45	genere		27	prova
17	237	nome		45	grado	93	26	ambiente
18	226	operazione	75	44	facolta'		26	bottega
19	217	qualita'		44	simile		26	casa
20	195	azione	76	43	moneta		26	denaro
21	189	sostanza	77	42	liquido		26	deposito
22	185	elemento		42	mammifero		26	diminuzione
23	183	apparecchio		42	minerale		26	esame
24	160	modo		42	rumore		26	gara
25	144	macchina		42	specie		26	imbarcazione
26	142	quantita'		42	struttura		26	ornamento
27	136	malattia		42	terreno		26	posizione
28	134	stato	78	41	insetto		26	sede
29	131	uccello		41	lavoro	94	25	capo
30	126	condizione	79	40	congegno		25	caratteristica
31	120	movimento		40	passaggio		25	cerimonia
32	119	tendenza		40	rappresentazione		25	festa
33	118	arte		40	strato		25	forza
	118	colpo	80	39	arbusto		25	negozio
	118	unita'		39	arma		25	piano
34	114	parola		39	colore		25	risultato
	114	scienza		39	composto		25	stabilimento
35	113	donna		39	roccia		25	varieta'
	113	gruppo		39	sala		25	vento
	113	tipo	81	38	dignita'	95	24	atleta
36	107	arnese		38	presenza		24	errore
37	105	mancanza	82	37	scritto		24	funzione
38	102	punto	83	36	branca		24	giocatore
39	99	oggetto		36	cavita'		24	pelle
	99	periodo		36	locale	96	23	artista
40	91	sistema		36	mobile		23	autorita'
41	89	titolo		36	rapporto		23	canto
42	88	albero	84	35	indumento		23	danno
	88	discorso		35	libro		23	elenco
43	87	striscia		35	tutto		23	manifestazione
44	86	recipiente	85	34	asta		23	mollusco
45	85	studio		34	aumento		23	notizia
	85	studioso		34	dolce		23	regione
46	84	uomo		34	materiale		23	riparo
47	83	denominazione		34	superficie		23	strada
48	82	segno		34	zona		23	successione
49	81	frutto	86	33	accordo		23	testo
50	80	forma		33	aspetto	97	22	attore
	80	gioco		33	costruzione		22	barca
	80	spazio		33	pietra		22	classe
51	77	piccolo		33	somma		22	comportamento
52	76	dispositivo		33	territorio		22	esposizione
	76	numero	87	32	fabbrica		22	istituto
	76	ufficio		32	fatto		22	nota
53	74	tecnica		32	foglio		22	perdita
54	73	tessuto		32	sensazione		22	tubo
55	72	figura		32	veicolo		22	valore
	72	processo	88	31	cifra	98	21	abitudine
56	71	fenomeno		31	cura		21	bastone
	71	organo		31	desiderio		21	estensione
57	70	pesce		31	documento		21	estremita'
58	69	capacita'		31	pratica		21	frase
	69	pezzo		31	stanza		21	momento
	69	termine	89	30	alterazione		21	polvere
59	65	linea		30	espressione		21	quello
	65	situazione		30	formazione		21	suddivisione
60	64	metallo		30	immagine		21	sviluppo
	64	sentimento		30	materia		21	taglio
61	63	atteggiamento		30	ramo		21	vino
	63	disposizione		30	tratto		21	vita
62	62	attivita'	90	29	disciplina	99	20	associazione
63	61	operaio		29	distanza		20	autore
	61	ordine		29	giorno		20	dichiarazione
	61	tempo		29	involucro		20	difetto
64	59	misura		29	scrittore		20	idea
65	58	componimento		29	unione		20	membro
	58	dottrina	91	28	bevanda		20	principio
66	57	edificio		28	famiglia		20	relazione
	57	opera		28	massa		20	riunione
	57	serie		28	metodo		20	sacerdote
67	56	nativo		28	mezzo		20	soluzione
	56	raccolta		28	senso		20	spettacolo

APPENDIX D

- Cibo 0 (1)
  - Ambrosia 1 (0)
  - Balsamo 0 (2)
  - Becchime 0 (0)
  - Boccone 0 (2)
    - Gnocco 0 (0)
    - Bocconcino 0 (1)
    - Leccornia 0 (0)
  - Bolo 0 (0)
  - Brodo 0 (0)
    - Consumato 0 (0)
    - Gelatina 0 (1)
  - Conserva 1 (2)
    - Confettura 0 (0)
    - Marmellata 0 (0)
  - Crema 0 (3)
    - Zabaione 0 (0)
  - Delicatezza 0 (3)
  - Esca 0 (1)
  - Ghiottoneria 0 (2)
  - Manna 1 (1)
  - Polenta 0 (0)
  - Polpetta 0 (2)
    - Granatina 0 (2)
  - Umido 0 (2)
  - Veleno 0 (2)
  - Vivanda 0 (0)
    - Crema 0 (1)
    - Dolce 0 (1)
      - Africano 0 (2)
      - Babà 0 (0)
      - Bavarese 0 (1)
      - Bignè 0 (0)
      - Brigidino 0 (0)
      - Cannoncino 0 (0)
      - Cassata 0 (2)
      - Chiacchiera 0 (4)
      - Chifel 0 (0)
      - Ciambella 0 (1)
      - Colomba 0 (2)
      - Cornetto 0 (3)
      - Croccante 0 (0)
      - Crostata 0 (0)
      - Diplomatico 0 (2)
      - Dolciume 0 (2)
      - Focaccia 0 (2)
        - Pizza 0 (0)
      - Frappe 0 (0)
      - Fritella 0 (0)
      - Gelato 0 (0)
        - Cassata 0 (1)
        - Granita 0 (0)
      - Mandorlato 0 (0)
      - Maritozzo 0 (0)
      - Meringa 0 (0)
      - Millefoglie 1 (0)
      - Pandoro 0 (0)
      - Panettone 0 (0)
      - Panforte 0 (0)
      - Panpepato 0 (0)
      - Semifreddo 0 (0)
      - Sfogliata 2 (0)
      - Strudel 0 (0)
      - Tortello 0 (2)
    - Farinata 0 (0)
    - Fonduta 0 (0)
    - Frittata 0 (0)
    - Frittura 0 (2)
    - Migliaccio 0 (1)
    - Pietanza 0 (0)
      - Giardiniera 0 (4)
      - Lesso 0 (0)
      - Manicaretto 0 (0)
      - Medaglione 0 (5)
      - Parmigiana 0 (0)
      - Peperonata 0 (0)
      - Polpetta 0 (1)
      - Granatina 0 (2)
      - Sformato 0 (0)
      - Spezzatino 0 (0)
        - Fricassea 0 (0)
        - Salmì 0 (0)
      - Stracotto 0 (0)
      - Stufato 0 (0)
    - San Giuseppe 0 (2)
    - Spiedino 0 (0)
    - Timballo 0 (0)
    - Tortino 0 (0)